一本临床医生看得懂、用得上的统计书

Step-by-step Examples of Data Analysis and Statistical Graphs in Clinical Researches

临床医学

研究中的统计分析和图形表达实例详解

EpiData+SPSS+GraphPad

主审◎马骏

主编◎周登远 崔壮 焦振山

军事医学科学出版社

VISIT…

LANZAROTE Caliente.com

◆EpiData 数据录入
◆SPSS18中文版 统计分析
◆GraphPad Prism 5 统计图绘制

如何使用本书：

◆如果你明确采用何种统计方法，直接翻到相应章节即可。

◆如果你不太明确采用何种统计方法，先翻到“第二章　统计方法的选择”，按图索引，找到相应章节。

◆如果你时间充裕，通读本书，定能构建你的统计思维，打造你的科研核心竞争力。

ISBN 978-7-80245-711-9

定价：43.00元

临床医学研究中的统计分析和图形表达实例详解

（EpiData + SPSS + GraphPad）

主　审　马　骏（天津医科大学公共卫生学院）

主　编　周登远（天津市中医药研究院）

　　　　崔　壮（天津医科大学公共卫生学院）

　　　　焦振山（天津市中医药研究院）

副主编　贾真琳（天津医科大学公共卫生学院）

　　　　赵光宇（军事医学科学院微生物流行病研究所）

　　　　颜利求（河北医科大学沧州临床医学院）

　　　　雷铭德（天津医科大学第二医院泌尿外科研究所）

编　者　（以姓氏拼音排序）

　　　　李长平　（天津医科大学公共卫生学院）

　　　　刘学慧　（河北医科大学公共卫生学院）

　　　　于　婷　（军事医学科学院微生物流行病研究所）

　　　　郑　纺　（天津中医药大学）

　　　　张　磊　（武警医学院）

　　　　赵林胜　（天津市儿童医院）

军事医学科学出版社

·北　京·

内容提要

本书分为预备篇、统计分析篇和统计绘图篇。预备篇介绍了统计学基本知识、统计方法的选择和如何采用 EpiData 或 SPSS 建立数据文件；统计分析篇以 SPSS 18 中文版为介绍对象，实例解说了计量资料、计数资料、生存资料和诊断试验中的统计分析，涵盖了临床医学科研中 90% 以上的数据分析内容；统计绘图篇则在国内首次介绍如何采用 GraphPad Prism 绘制出版级统计图，以统计方法为框架，以实例详解的方式，简单明了。本书以“实用”为出发点，力图在最短时间最大限度地满足临床医学科研中的实际需求。本书见解独到、语言幽默、简单实用，适合临床医生、研究生和高年级本科生，同时对基础医学科研工作者也有极大的参考价值。

图书在版编目(CIP)数据

临床医学研究中的统计分析和图形表达实例详解/周登远，崔壮，焦振山主编.
-北京：军事医学科学出版社，2011.2
ISBN 978 - 7 - 80245 - 711 - 9

Ⅰ.①临… Ⅱ.①周… ②崔… ③焦… Ⅲ.①医学统计 - 统计分析 - 应用软件 Ⅳ.①R195.1 - 39

中国版本图书馆 CIP 数据核字(2011)第 018222 号

出　版：军事医学科学出版社
地　址：北京市海淀区太平路 27 号
邮　编：100850
联系电话：发行部：(010)66931051，66931049，63827166
编辑部：(010)66931039，66931127，66931038
86702759，86703183
传　真：(010)63801284
网　址：http://www.mmsp.cn
印　装：北京市顺义兴华印刷厂
发　行：新华书店

开　本：787mm × 1092mm　1/16
印　张：17.5
字　数：427 千字
版　次：2011 年 5 月第 1 版
印　次：2011 年 5 月第 1 次
定　价：43.00 元

本社图书凡缺、损、倒、脱页者，本社发行部负责调换

出版说明

“统计有用,统计难学”几乎成了所有临床研究人员共同的心声,大家对统计的理解还停留在那些晦涩难懂的数学公式,以及以编程为基础的SAS软件。虽然从本科开始就开设了统计学课程,但最后临床医生能独立完成数据分析和处理成了一种奢望,慢慢地甚至产生了统计恐惧症。该书的作者均是常年从事临床数据分析的一线统计专家,多年从事临床科研咨询、统计教学和稿件审阅等工作,结合工作,越发觉得需要编写一本临床医生看得懂、用得上的统计实用书籍,本书具有如下特点:

一、选择 EpiData 3.1 录入数据,SPSS 18 中文版统计分析,GraphPad Prism 5 绘制图形

人生的智慧在于选择,作为一名非统计专业人员,选择软件需要兼顾易用性和权威性,而各医科大学在临床研究生中教授SAS的使用,如同教一个不会开车的人参加F1方程赛,其结果可想而知了。而同样权威的SPSS 18中文版,以全中文的图形界面,使独立完成统计分析变得如此简单。而采用GraphPad Prism通过简单几步即可完成出版级的统计图绘制,EpiData对于大型问卷调查录入具有无法比拟的优势。综合发挥各软件的优势,让它们成为你科研生涯的忠实伴侣。

二、以数据类型为框架,让你明明白白选择统计方法

传统的统计学教科书是以统计理论为框架,从分布开始谈起,不断深入;而统计软件教程则往往以软件界面为框架,逐个菜单栏进行介绍,但这两种框架在实际运用中会产生一些问题,一个研究中必定存在这些数据资料“能进行哪些统计分析”,以及如果要回答研究者提出的研究假设“需要采用何种统计分析”,所以“能做什么”和“想做什么”成为研究过程中最关心的问题。本书以数据资料(能做什么)为框架,在各种资料中,又以研究假设(想做什么)作为分类标准来介绍统计方法,详尽介绍了计量资料、计数资料、生存资料和诊断试验这四类临床科研中最常用数据资料的统计方法,而撇弃了医学临床科研中很少使用的聚类分析、时间序列分析等方法。中国有句古话“有所为,有所不为”,让一个临床医生掌握经济预测方法和市场调查分析意义并不很大。本书还单独列出“第二章　统计方法的选择”,以帮助大家更好更迅速地选用恰当的统计方法。

三、各节单独成章，以线性流程的讲解方式让你迅速掌握统计应用

书中大部分例子取自方积乾教授主编的卫生部统编教材《卫生统计学》(第6版)，熟悉的例题，通过SPSS软件再次重现。每节均由方法原理－分析示例－研究假设－数据录入－操作流程－结果解释－注意事项这样7个部分组成，你只要按书中提示一步一步完成即可掌握该统计方法。

四、以统计方法为框架介绍统计图绘制

以往书籍介绍统计图形均就图论图，逐一介绍直方图、条图、饼图和线图等的绘制，但在实际应用中对数据采用何种统计图形表达却是个难题，该书则以统计方法为框架，将统计分析和统计图形联系起来，让大家迅速选用合适的统计图形。

五、各节单独成章，以线性流程的讲解方式让你迅速掌握图形绘制

本书在国内首次对GraphPad Prism软件应用进行系统介绍，所有章节均由示例－该示例的图形表达－图形类型的选择－数据录入－坐标轴和图形调整－文字修饰6个部分组成，让你掌握发表级图形的绘制。

本书以"一切为了应用"为出发点，深刻把握临床医生在科研中的实际需求，帮助大家迅速有效地掌握统计知识，增强科研核心竞争力。

本书数据文件 www. mmsp. cn

作者信箱：周登远 zhoudengyuan@ hotmail. com；崔壮 nicholastricle@ hotmail. com
作者博客：www. zhoudenyuan. com

周登远　崔壮　焦振山

目 录

上篇　预备篇

第一章　统计学的基本概念 …… (3)
第二章　统计方法的选择 …… (6)
　第一节　计量资料的统计方法选择 …… (6)
　第二节　计数资料的统计方法选择 …… (8)
　第三节　生存资料的统计方法选择 …… (10)
　第四节　诊断试验资料的统计方法选择 …… (11)
　第五节　两个大家容易迷惑的问题 …… (11)
第三章　数据文件的建立 …… (12)
　第一节　用 SPSS 建立数据文件 …… (12)
　第二节　EpiData 数据录入 …… (17)

中篇　统计分析篇(SPSS 18 中文版)

第四章　*t* 检验 …… (25)
　第一节　单样本 *t* 检验 …… (25)
　第二节　配对 *t* 检验 …… (28)
　第三节　成组 *t* 检验 …… (31)
第五章　方差分析 …… (35)
　第一节　完全随机设计资料的方差分析 …… (36)
　第二节　随机区组设计资料的方差分析 …… (42)
　第三节　析因设计资料的方差分析 …… (50)
　第四节　重复测量资料的方差分析 …… (54)
第六章　秩和检验 …… (65)
　第一节　单样本秩和检验 …… (65)
　第二节　配对样本秩和检验 …… (69)
　第三节　两组独立样本秩和检验 …… (74)
　第四节　多组独立样本秩和检验 …… (79)

第七章　相关分析 ………………………………………………………………………… (85)
第一节　线性相关 ………………………………………………………………………… (85)
第二节　秩相关 ………………………………………………………………………… (90)
第八章　线性回归分析 ………………………………………………………………… (94)
第一节　简单线性回归分析 …………………………………………………………… (95)
第二节　多重线性回归分析 …………………………………………………………… (100)
第九章　四格表卡方检验 ……………………………………………………………… (107)
第一节　一般四格表卡方检验 ………………………………………………………… (107)
第二节　配对四格表卡方检验 ………………………………………………………… (113)
第十章　列联表分析 …………………………………………………………………… (118)
第一节　双向无序的列联表分析 ……………………………………………………… (118)
第二节　单向有序的列联表分析 ……………………………………………………… (124)
第三节　双向有序且属性不同的列联表分析 ………………………………………… (129)
第四节　双向有序且属性相同的列联表分析 ………………………………………… (135)
第十一章　Logistic 回归 ……………………………………………………………… (140)
第一节　非条件 Logistic 回归 ………………………………………………………… (140)
第二节　条件 Logistic 回归 …………………………………………………………… (148)
第十二章　生存分析 …………………………………………………………………… (157)
第一节　寿命表法 ……………………………………………………………………… (159)
第二节　单因素生存曲线比较(Kaplan-Meier 法)…………………………………… (162)
第三节　多因素生存分析(Cox 回归分析) ………………………………………… (169)
第十三章　诊断试验的统计分析 ……………………………………………………… (178)
第一节　准确性检验(ROC 曲线) …………………………………………………… (178)
第二节　一致性检验(Bland-Altman 图) …………………………………………… (184)

下篇　统计绘图篇(GraphPad Prism 5)

第十四章　统计图基本知识 …………………………………………………………… (195)
第十五章　GraphPad Prism 5 绘图界面介绍 ………………………………………… (199)
第十六章　各种统计方法所对应的统计图绘制 ……………………………………… (205)
第一节　配对 t 检验的图形绘制……………………………………………………… (205)
第二节　成组 t 检验的图形绘制……………………………………………………… (208)
第三节　完全随机设计资料方差分析的图形绘制 …………………………………… (211)
第四节　析因设计资料方差分析的图形绘制 ………………………………………… (215)
第五节　重复测量资料方差分析的图形绘制 ………………………………………… (218)
第六节　两组独立样本秩和检验的图形绘制 ………………………………………… (222)
第七节　多组独立样本秩和检验的图形绘制 ………………………………………… (225)

第八节　简单线性回归和线性相关的图形绘制 …………………………………（229）
第九节　列联表分析的图形绘制 ……………………………………………（233）
第十节　生存分析的图形绘制 ………………………………………………（238）
第十七章　统计图的排版与导出 ………………………………………………（245）

附录一　EpiData,SPSS,GraphPad Prism 介绍 …………………………………（253）
附录二　临床研究国际论文撰写指南 ……………………………………………（258）
附录三　参考文献 …………………………………………………………………（267）

上篇

预备篇

>> 第一章　统计学的基本概念

一、样本与总体

1. 总体(population)　是根据观察目的而确定的同质观察单位的全体,即同质的所有观察单位某种变量值的集合。

2. 样本(sample)　样本是总体中随机抽取的部分观察单位的实测值的集合。

科学研究一般是通过样本来推断总体特征,其做法是从研究总体中抽取少量有代表性的个体,称为抽样(sampling),对这些个体组成的样本进行深入的观察与测量,获取数据(data),利用统计知识,透过样本数据对研究总体的规律进行推断(inference)。

二、变异与同质

1. 变异(variation)　同质个体同指标之间的差异称为变异。

2. 同质(homogeneity)　指研究事物现象存在的共性。它是统计研究的基础,是资料整理和分析的前提。

三、变量的分类

变量(variable):总体中个体特征总是通过一个或多个变量来描述,变异性的存在决定了我们要处理的是变量。本书把变量分为定性(qualitative)和定量(quantitative)两种。

1. 定性变量又分为分类变量和有序变量(等级变量)

(1)分类变量(categorical variable):又称名义变量(nominative variable),例如职业是一个分类变量,其可能的"取值"不是数字,而是工、农、商、学、兵等,这些成为分类变量的水平(level),为便于输入计算机,一般采用代码(code)1、2、3、4、5 等来表示各水平。最简单也是最常用的变量为二分类变量(binary variable),如性别(男、女)、疾病(有、无)和结局(生、死)等。

(2)有序变量(ordinal variable):指分类变量种种可能的"取值"中自然地存在着次序。例如,问卷调查中常问对某件事情的满意程度,给出了 5 个答案:极不满意、有点满意、中度满意、很满意、极满意。有些临床体检或实验室检验常用 -、±、+、++ 和 +++ 来表示测量结果。

2. 定量变量又分为离散变量和连续变量

(1)离散变量(discrete variable):离散型变量只能取整数值,如一个月中手术病人数,一年里的新生儿数。

(2)连续变量(continuous variable):连续型变量可以取实数轴上的任何数值。有些变量的数值由测量而得到,它们大多属于连续变量,如血压、身高、体重等。而有一些测量值,如红细胞数,虽然以"个"为单位时只能取整数值,但当数值很大而以"千"或"万"为单位时,又可以

取小数值，所以通常把这些变量也称为连续变量。

有时为了数据分析的方便，人们将一种类型的变量转化为另一种类型，但变量只能由“高级”向“低级”转化（定量变量—有序变量—分类变量—二分类变量）。

四、频率与概率

（一）频率（frequency）

指在相同条件下，进行有限 n 次重复试验，某随机事件 A 发生的次数与 n 次试验的比值，频率是个变数，随样本变化而改变。

（二）概率（probability）

是描述随机事件 A 发生的可能性大小的度量，概率是一个定值。假设在相同的条件下，独立进行 n 次重复试验，随着 n 逐渐增加，频率摆动的幅度越来越小，则该事件 A 为随机事件，其频率可作为概率的估计值。

五、误差的分类

误差（error）可以分为随机误差和非随机误差。

1. 随机误差又分为抽样误差和随机测量误差

（1）抽样误差：由于产生的根本原因是生物个体的变异性，故抽样误差的分布具有规律性。

（2）随机测量误差：对同一观察单位某项指标在同一条件下进行反复测量所产生的误差。

2. 非随机误差又分为系统误差和过失误差

（1）系统误差：可产生于设计人员、调查者或调查对象，也可由于考虑不当、汇总计算有误等造成，一般带有倾向性，其产生原因复杂，贯穿于研究全过程并对研究结果有影响，很难用统计方法评价它的影响。

（2）过失误差：是错误，一般应杜绝。

六、统计分析的流程

（一）根据临床实践，提出研究问题，进行科研设计

在医学实践过程中提出科研问题，然后围绕提出的研究问题，制订研究方案，统计分析人员应当从设计阶段就参与研究项目，而不是临床医生获得数据之后，才想到统计分析。医学研究一般有干预性研究（intervention study）和观察性研究（observational study）两种。医学干预性研究是人们通过规定对象的准入条件（entry criteria）、随机化、重复、匹配（match）以及盲法（blinding）等措施来控制主要的混杂因素。公共卫生方面的研究大多属于观察性研究，这类研究不可能人为地控制许多混杂因素，人们能做的主要是观察已经或将要发生的事情。对于混杂因素的处理办法是精心设计抽样方法、无误地记录可能有用的信息。

（二）进行科学研究，分析清楚资料的性质，并分解出观察与变量

研究方案出来后，需要严格按照研究方案进行，资料大体上可以分为计量资料和计数资料，计量资料指测定每个观察单位的某项指标量的大小所得的资料；而将观察单位按照某种属性或类别分组计数，所得各组观察单位数称为计数资料。分清楚资料类型后，需要将资料分解成观察与变量。变量在临床上称为指标，是指具有相同属性的测量值的集合；而观察是指对同

一观察对象的不同属性的集合，就构成了一条观察；观察与变量结合起来就能准确地描述二维空间的物体特征。区分资料为计量资料还是计数资料，然后将资料分解为观察与变量，这是资料分析的基础。

（三）结合以上两点，罗列出能够回答该问题的可选统计方法

根据研究目的和资料性质，选用相应的统计学方法，如计量资料中的 t 检验、方差分析、线性回归分析等，计数资料中的卡方检验、Logistic 回归分析等。不同的资料和研究目的有多种可供选择的统计方法；而多种不同的统计方法可以对应多种资料类型，回答多种问题，如秩和检验能处理不符合方差分析条件的计量资料，也可以分析等级资料。

（四）选用统计软件，尝试着进行相关的统计分析

大家经常认为，只要形成了数据表格，选对了统计方法，用软件一操作，就万事大吉了。其实，事情没有这么简单，统计分析是一个反复的过程，是一个系统工程，需要进行预分析、正式分析等，如一份计量资料，我们首先考虑进行方差分析，但是分析过程中发现方差不齐，我们可以改做秩和检验。如在 Logistic 回归过程中，可以选用全部进入法和逐步回归法，也可以两者均尝试一下，然后比较两种方法所得出结果的差异，再根据专业知识和分析目的，作出判断，可见统计分析不是一锤定音、一成不变的过程，而是不断尝试、不断思考、不断判断的过程。

（五）评估统计结果，结合专业来回答提出的研究问题

从统计结论到专业结论，大家都需要特别慎重，不可妄加推断，任意发挥。

>> 第二章　统计方法的选择

统计方法的选择既是一个科学问题，也是一个艺术问题，同样的数据可以采用不同的统计方法，而不同的数据也可以采用同一种统计方法，因此我们需要在统计方法的选择中把握其根本：数据资料的性质决定“能做什么”，而研究设计或研究目的则决定了“想做什么”。本书将数据按性质分为计量资料和计数资料，对于两种资料又根据不同的研究设计或研究目的，讨论其具体的统计方法。而生存分析和诊断试验两章，则属于特定的分析方法，其数据特征和分析目的有其固定的特征，所以单独列出进行讨论。

需要特别说明的是，计量资料和计数资料只是对资料的一种通俗叫法，并不是计量资料中没有定性变量，也不是计数资料中没有定量变量。我们测量和感兴趣的指标为定量变量（如每组病例的血压值）或定性变量（如每组中治愈的人数），我们就称之为计量资料或计数资料。下面就计量资料、计数资料、生存资料和诊断试验的统计方法的选择进行讲解。

第一节　计量资料的统计方法选择

计量资料的统计分析按照研究设计或研究目的的不同，可以分为三类：

成组设计：其目的在于比较各组所代表的总体之间均数或中位数的差别，包括 t 检验、方差分析和秩和检验三种类型。

相关分析：其目的在于研究两个变量之间联系的密切程度，又可分为线性相关和秩相关。

因果联系：其目的在于探讨自变量和因变量之间的因果关系，因变量又称为结果变量，通常为身高、血压等连续变量；自变量又称解释变量，自变量可以为连续变量、等级变量和分类变量，其分析方法称为线性回归分析。

一、成组设计

1. 简单成组设计的统计方法选择

单组或两组计量资料的统计方法

设计名称	前提条件是否满足及假设检验方法的选择	
	满足	不满足
单组设计	单样本 t 检验	单样本秩和检验
配对设计	配对 t 检验	配对样本秩和检验
成组设计	成组 t 检验	两组独立样本秩和检验

上表统计方法是成组设计中最简单的统计方法，也是应用非常广泛的统计方法。t 检验中要求数据来自正态总体，如果这一前提条件不满足，则采用对应的非参数检验（秩和检验）。后面介绍的方差分析和秩和检验可以视为该表的扩展。

2. *复杂成组设计的统计方法选择* 在谈到复杂成组设计之前，有三个概念必须弄清楚：

（1）因素（factor）：指对测量结果可能有影响的变量，一般来说，因素不止一个水平，而分析的目的就是比较同一因素内各水平之间测量结果是否相同。在方差分析中，一般有一个或多个因素。

（2）水平（level）：因素的不同取值称为水平，例如，因素“性别”有男、女两个水平，需要注意的是有时水平是人为划分的，比如身高被分为高、中、低三个水平。

（3）交互作用（interaction）：如果一个因素的效应大小在另一个因素不同水平下明显不同，则称为两因素间存在交互作用。当存在交互作用时，单纯研究某个因素的作用是没有意义的，必须在另一个因素的不同水平下研究该因素的作用大小。有时两因素之间的交互作用无法测量，如随机区组设计的方差分析。

根据因素、水平和交互作用的不同，可以将方差分析分为如下几类：

方差分析类型

方差分析类型	因素	水平	交互作用
完全随机设计	1	>2	无
随机区组设计	2	>2	无
析因设计	2	≥2	有

（1）完全随机设计可以视为成组 t 检验的扩展，两者均为一个因素，但是成组 t 检验中水平数为 2，而完全随机设计方差分析中水平数大于 2。由于只有一个因素，故不存在交互作用。

（2）随机区组设计存在两个因素，但两个因素的地位并不相同，如考察不同饲料剂量对大白鼠体重的影响，饲料是研究因素，而大白鼠窝别为区组因素，区组因素是为了消除混杂因素而引入的，从研究设计上就要求饲料和窝别之间不存在交互作用，否则该设计不合理。

（3）析因设计存在两个因素，且两个因素地位相同，如考察缝合方法和缝合时间对于大白兔神经损伤后愈合的影响（测量指标为轴突通过率，为计量资料），分析时就应当考虑缝合方法和缝合时间是否存在交互作用。

复杂成组设计的统计方法选择

设计名称	前提条件是否满足及假设检验方法的选择	
	满足	不满足
完全随机设计	完全随机设计的方差分析	Kruskal-Wallis 秩和检验
随机区组设计	随机区组设计的方差分析	Friedman 秩和检验
析因设计	析因设计的方差分析	非参数较少，进行数据变换
重复测量设计	重复测量设计的方差分析	

此处需要注意方差分析的适用条件：

(1)独立性：要求各样本为相互独立的随机样本，才能保证变异的可加性(可分解性)。

(2)正态性：即所有观察值均从正态总体中抽样得出。

(3)方差齐：指假设总的模型无意义时方差齐。

以上条件中，独立性要求最严，其次为正态性和方差齐性。在重复测量设计中，由于各次测量违反了独立性原则，所以采用特殊的重复测量的方差分析。

二、相关分析

(1)线性相关：两个随机变量之间的联系，即适用于二元正态分布的资料，常用 Pearson 相关系数表示。

(2)秩相关：对于不服从双变量正态分布的资料，还有总体分布未知的资料和原始数据用等级表示的资料，常用 Spearman 秩相关系数表示。

三、因果联系

(1)简单线性回归：因变量(结果变量)为连续变量，自变量(解释变量)也只有一个连续变量。

(2)多重线性回归：因变量(结果变量)为连续变量，自变量(解释变量)有多个变量，可以为连续变量、等级变量和分类变量。统计软件只能处理连续变量，分类变量可以转换为哑变量处理，等级变量则可以按连续变量或哑变量处理。

第二节　计数资料的统计方法选择

R×C 表：包括四格表和列联表，是计数资料中最常见的一种类型。

因果联系：其目的在于探讨自变量和因变量之间的因果关系，因变量又称为结果变量，通常为二分类变量、多分类变量和等级变量，本书只谈到最常用的二分类变量；自变量又称解释变量，可以有多个自变量，自变量可以为连续变量、等级变量和分类变量，其分析方法称为 Logistic 回归。

一、R×C 表

四格表统计分析

项目	统计方法
一般四格表	χ^2检验、Fisher 精确检验
配对四格表	McNemar 检验、Kappa 检验

列联表的统计分析方法

变量的统计性质及其专业属性	列联表分类	可选用的统计方法
X,Y 皆为分类变量且属性不同	双向无序表	χ^2 检验,Fisher 精确检验
X 为分类变量,Y 为有序变量	单向有序表	秩和检验,Ridit 分析,有序变量的 Logistic 回归
X,Y 皆为有序变量且属性不同	双向有序表	关心组间差别,按单向有序的列联表处理
		是否相关,用 Spearman 秩相关或典型相关分析
		是否存在直线变化,用线性趋势检验
X,Y 皆为有序变量且属性相同	双向有序表	一致性检验(即 Kappa 检验)

注:几乎所有列联表资料都可以用对数线性模型或 Logistic 回归模型来分析

列联表的类别判断是大家头痛的地方,大家看看以下的四个例子就会一目了然。

1. *双向无序列联表*　研究问题:两种血型划分是否独立,即 ABO 血型与 MN 血型类别的构成比有无差别。

某地 6094 人 ABO 及 MN 血型分布情况

ABO 血型	MN 血型			合计
	M	N	MN	
O	431	490	902	1832
A	388	410	800	1598
B	495	587	950	2032
AB	137	179	325	641
合计	1452	1666	6094	6094

2. *单向有序列联表*　研究问题:三种药物的疗效有无差别。

3 种药物疗效的观察结果

疗效	药　　物			合计
	A	B	C	
治愈	15	4	1	20
显效	49	9	15	73
好转	31	50	45	126
无效	5	22	24	51
合计	100	85	85	270

3. 双向有序且属性不同的列联表

研究问题1:三个年龄组的晶体混浊程度是否相同。

研究问题2:年龄与晶体混浊程度是否相关。

研究问题3:随着年龄的增加,晶体混浊程度是否呈现某种变化趋势。

眼晶状体混浊度与年龄的关系

晶状体混浊程度	年龄			合计
	20 -	30 -	40 -	
+	215	131	148	494
+ +	67	101	128	296
+ + +	44	63	132	239
合计	326	295	408	1029

4. 双向有序且属性相同的列联表　研究问题:两种测量方法的一致性到底有多大。

两种方法检查室壁收缩运动的符合情况

对比法测定结果	核素法			合计
	正常	减弱	异常	
正常	58	2	3	63
减弱	1	42	7	50
异常	8	9	17	34
合计	67	53	27	147

二、因果联系

本书只探讨二分类 Logistic 回归,如果需要进行多分类 Logistic 回归或等级 Logistic 回归,请参阅相关书籍。条件 Logistic 回归和非条件 Logistic 回归中因变量(结果变量)均为二分类变量,自变量(解释变量)有多个变量,可以为连续变量、等级变量和分类变量。统计软件只能处理连续变量,分类变量可以转换为哑变量处理,等级变量可以按连续变量或哑变量处理。

1. 非条件 Logistic 回归　指病例组和对照组采用非匹配设计。

2. 条件 Logistic 回归　指为病例组的每个研究对象匹配一个同样特征的未患病者,作为该病例的对照,又称 1:1 配对设计,该设计能提高研究效能。

第三节　生存资料的统计方法选择

生存数据是一类特殊的数据,其数据类型(能做什么)和研究目的(想做什么)都是固定的,是一个货真价实的二合一产品。生存数据同时考虑了结局以及达到终点所需要的时间,并

且能够处理失访的研究对象(又称截尾值),这是生存分析最大的特点。按照研究目的和资料的不同又可分为以下三类:

1. *生存描述*　采用寿命表法,适用于分段统计的资料,即将整个观察时间划分成很多小的时间段,然后统计各时间段内发生的终点事件(如死亡)和失访的数目。

2. *成组设计*　采用 Kaplan-Meier 法,资料为各研究对象出现终点事件的具体时间或者发生失访的具体时间,其研究目的为比较两组或多组的研究对象生存曲线,一般需要同时绘制生存曲线图。

3. *因果联系*　采用 Cox 回归,其因变量(结果变量)为观察时间以及到达该观察时间时终点事件是否发生(如死亡或失访),而自变量(解释变量)可有多个变量,可以为连续变量、等级变量和分类变量。统计软件只能处理连续变量,分类变量可以转换为哑变量处理,等级变量可以按连续变量或哑变量处理。

第四节　诊断试验资料的统计方法选择

诊断试验是指对疾病进行诊断和鉴别诊断的试验方法,按照其目的一般可分为以下两类:

1. *准确度检验*　即常见的 ROC 曲线,其资料要求为连续变量或等级变量,在经过金标准判断后,计算其灵敏度和特异度等指标,可以探讨两种试验诊断方法的检测效果。

2. *一致性检验*　即绘制 Bland-Altman 图,其资料要求两种检测方法获得的检测结果为连续变量,用图表示两种检测方法的一致性。

第五节　两个大家容易迷惑的问题

(一)为什么本书没有讲解统计描述部分

一般而言,在科研资料分析中,很少出现单纯的统计描述,而是统计描述和统计推断同时进行,先对资料进行描述,然后对资料进行推断。SPSS 统计分析结果也同时包含了统计描述和统计推断,所以该书没有单独列出统计描述部分,各种数据的描述也在相应的统计分析部分给出。

(二)多重线性回归、Logistic 回归和 Cox 回归之间是否存在某种联系

三者均属于回归分析,其目的在于探讨多个自变量(解释变量)对因变量(结果变量)的影响,并且自变量具有共同属性。其联系见下表:

回归分析	因变量(结果变量)	自变量(解释变量)
多重线性回归	连续变量	存在多个变量,可以为连续变量、等级变量和
Logistic 回归	二分类变量、多分类变量或等级变量	分类变量,分类变量转换为哑变量进行处理,
Cox 回归	时间变量和二分类变量	等级变量按连续变量或哑变量进行处理

>> 第三章　数据文件的建立

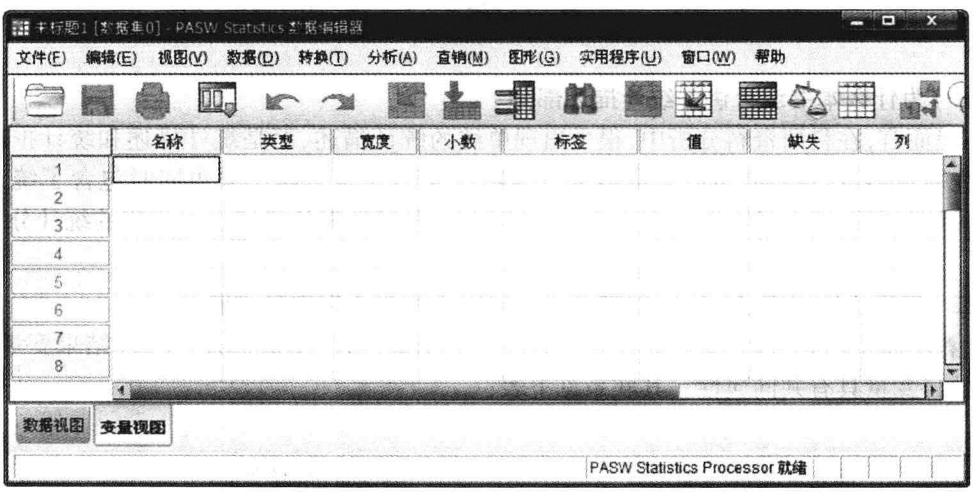
建立数据文件，将研究中所获得的数据转变为SPSS统计软件可以分析的.sav格式数据库，这是统计分析的第一步，但不巧的是，很多临床研究人员在这里就卡壳了。本章分别介绍了用SPSS和EpiData 3.1软件建立数据文件，简单的、少量的数据就用SPSS建立，复杂的、大量的数据就用EpiData建立。在以后章节的每一个统计方法示例中，均会详细给出如何用SPSS建立数据文件，有这样手把手的帮助，一切都变得简单了。

第一节　用SPSS建立数据文件

一、SPSS数据录入的两个窗口

打开SPSS软件，出现了SPSS主窗口，主窗口又分为变量视图窗口和数据视图窗口。

(一)变量视图窗口



数据录入的第一步是建立变量，并对其进行设定，这一切均在变量视图窗口进行，其常用的选项包括名称、类型、标签、值这样几项，其中名称和类型为变量的基本属性，而标签和值只会使结果显示更直观，下面分别介绍：

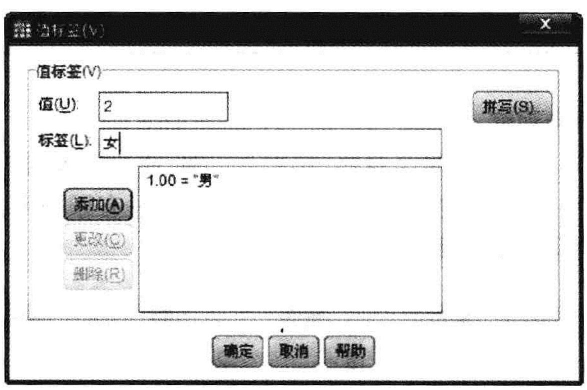
名称：即变量名，这项在变量设定中最为重要，是变量唯一识别号，相当于变量的身份证。以前版本SPSS一般要求为英文字母和数字，现在也可以采用中文，但我们习惯于还是采用英文字母，如“年龄”变量名为age。

类型:变量类型有数值、逗号、点、科学计数法、日期、美元、设定货币和字符串这样几项,但常用的为数值、字符串和日期三项,默认为数值,一般设为数值即可。如果原始数据为中文字符,如"甲、乙、丙"等,或"+、++、+++",则需要将类型设为字符串。日期类型一般在生存分析时采用,实际上,SPSS 中日期型变量的存储是该时间与 1582 年 10 月 14 日零点相差的秒数,如 1582 年 10 月 15 日存储的是 60×60×24=86 400,大家需要注意的是,SPSS 日期型变量输入格式有 dd-mmm-yyyy、dd-mmm-yy、mm/dd/yyyy、mm/dd/yy 多种,你需要选择与要输入的原始表格最相近的格式,这样就能减少录入错误。

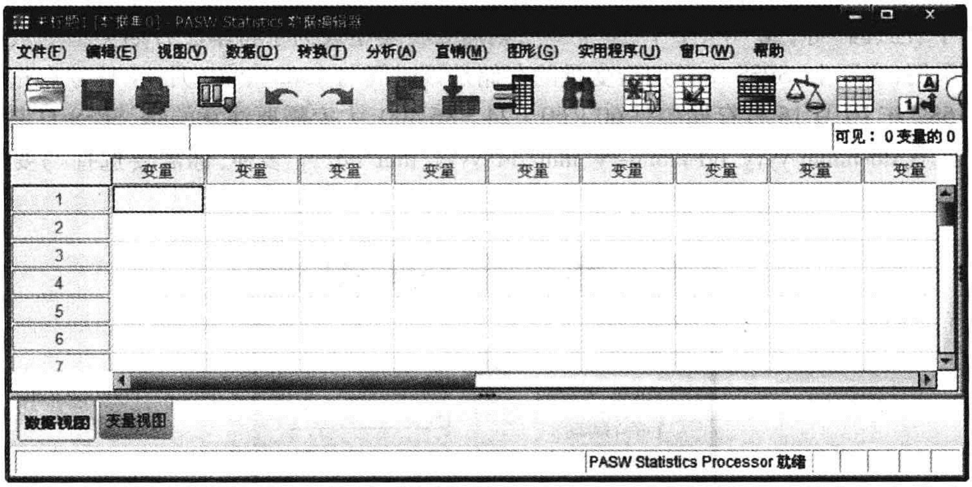


标签:也就是给变量取一个别名,一般用中文来表示,如变量名为 name,则标签则为"姓名",以便于识别。

值:表示变量下各水平的具体含义,如变量"sex",类型为"数值",标签为"性别",其变量有两个水平 1 和 2,其中 1 代表"男",2 代表"女",见下图:



（二）数据视图窗口



在变量视图窗口设定好了变量，现在可以在数据视图窗口输入数据，纵向为变量，横向为观察，如设置了变量 id，sex，age，第一条观察可能为 1，1，35，横向与纵向联合起来形成二维数据表格。

二、计量资料的数据文件建立

示例

分别测得 5 例老年慢性支气管炎病人及 5 例正常人的尿类固醇排出量（mg/dl）如下，试比较两组的均数有无差别。

病人	2.90	5.41	5.48	4.60	4.03
健康人	5.18	8.79	3.14	6.46	3.72

Step 1：**分解出变量和观察**

对于数据，需要首先分解出变量与观察，在这个步骤，就把很多同学给难倒了，如本例可将其分解出两个变量：组别 group（包括病人组和健康人组），还有尿类固醇排出量 x。

Step 2：**变量视图窗口设定变量**

	名称	类型	宽度	小数	标签	值
1	group	数值(N)	8	2	组别	{.0, 正常人}...
2	x	数值(N)	8	2	尿类固醇排出量	无

<u>名称</u> group <u>类型</u> 数值 <u>标签</u> 组别 <u>值</u> 0 =“正常人”1 =“病人”

<u>名称</u> x <u>类型</u> 数值 <u>标签</u> 尿类胆固醇排出量

Step 3:**数据视图窗口输入数值**

	group	x
1	1.00	2.90
2	1.00	5.41
3	1.00	5.48
4	1.00	4.60
5	1.00	4.03
6	.0	5.18
7	.0	8.79
8	.0	3.14
9	.0	6.46
10	.0	3.72

注意:不同的统计方法,对数据表中的变量和观察有不同的要求,如果为配对 t 检验,则要求两个变量为 patient(病例组),control(对照组),大家也不要担心,每个例子本书都给出了详细的数据录入步骤,大家只需要照猫画虎即可。

三、计数资料的数据文件建立

示例

两种药物治疗慢性咽炎,其治疗具体效果见下表:

两种药物治疗慢性咽炎疗效情况

药物(drug)	疗效(effect)		合计
	有效(effect = 1)	无效(effect = 0)	
A(drug = 1)	1	2	3
B(drug = 2)	4	3	7
合计	5	5	10

(一)方法 1

Step 1:**分解出变量和观察**

四格表是计数资料中最常见的数据形式,也是 R × C 表的基本格式,一般将其分解为治疗分组(drug), 疗效分组(effect),权重变量(weight)。

Step 2:**变量视图窗口设定变量**

	名称	类型	宽度	小数	标签	值
1	drug	数值(N)	8	0	药物	{1, A药}...
2	effect	数值(N)	8	0	疗效	{0, 无效}...
3	weight	数值(N)	8	0	权重	无

名称 drug　　类型 数值 标签 药物　值 1 = "A 药"2 = "B 药"
名称 effect　　类型 数值 标签 疗效　值 0 = "无效"　1 = "有效"
名称 weight　　类型 数值 标签 权重

Step 3:**数据视图窗口输入数值**

	drug	effect	weight
1	1	1	1
2	1	0	2
3	2	1	4
4	2	0	3

(二)方法 2

Step 1:**分解出变量和观察**

四格表是计数资料中最常见的数据形式,也是 R×C 表的基本格式,有些原始的记录只有两个变量治疗分组(drug),疗效分组(effect),没有权重变量(weight)这个值。

Step 2:**变量视图窗口设定变量**

	名称	类型	宽度	小数	标签	值
1	drug	数值(N)	8	0	药物	{1, A药}...
2	effect	数值(N)	8	0	疗效	{0, 无效}...

名称 drug　　类型 数值 标签 药物　值 1 = "A 药"　2 = "B 药"
名称 effect　　类型 数值 标签 疗效　值 0 = "无效"　1 = "有效"

Step 3:**数据视图窗口输入数值**

	drug	effect
1	1	1
2	1	0
3	1	0
4	2	1
5	2	1
6	2	1
7	2	1
8	2	0
9	2	0
10	2	0

(三)两种建立数据库方法的区别

从输入的角度讲,方法1要简单,也比较常见。

从使用的角度讲,在进行分析之前,方法1必须指定权重变量,即需要作如下操作:

数据(D)—加权个案(W) 频率变量(F):权重 weight



第二节　EpiData　数据录入

虽然用SPSS可建立简单的数据文件,但是在原始资料收集时,如果数据很多而且复杂,如流行病学调查,通常采用EpiData作为数据录入工具。本章以EpiData 3.1为介绍对象,通过一个简单的示例让大家掌握EpiData的使用方法。

下面是一个纸质的调查问卷表,调查员通过询问,将原始数据记录在调查问卷上,然后通过Epidata 3.1将问卷内容转化为电子数据库,以便通过SPSS等统计软件进行处理。

EpiData 调查问卷示例

1. 姓名 ______
2. 年龄 ____岁
3. 性别:男　女
4. 您现在是否患有糖尿病?

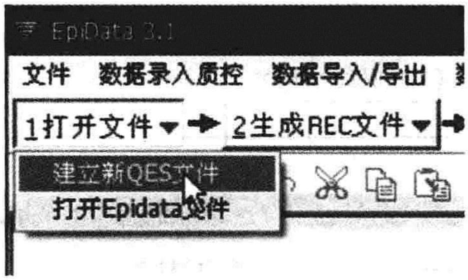
   ①是　　②否(跳转至第6题)
5. 您现在所选择的降糖方案为以下哪一种?

   ①只服用降糖药　②只注射胰岛素　③口服降糖药+注射胰岛素
6. 您现在是否患有高血压?

   ①是　②否

Step 1:**建立 QES 文件**

单击“1 打开文件”,出现下拉菜单,选择“建立新 QES 文件”,下方的灰色区域和工具栏均由灰色转变为激活状态。



我们就可以根据调查表的内容,在空白区域内定义数据,编辑 QES 文件。编写 QES 文件时要注意,QES 文件的内容与格式应尽量与调查表保持一致,这样能减少以后的录入错误。

编写 QES 文件时,每一个变量都由 3 部分信息组成:

1. 变量名　变量名一定要以英文字母开头,如实例中的 n1,n2 等。如果一个问题里有多个小问题,可再细分为 n1a,n1b……。

2. 变量标记　可用中文形式解释该变量。

3. 变量值　为需要录入的数据区域。根据变量的字段类型不同,变量值的定义方式也会发生相应变化。

文本类型用“________”定义;一个中文字占两个字符,一个英文字母占一个字符。

数字类型用“#”定义,一个#代表一个数字,如可能取值为个位数,则将变量定义为#,如可能取值为 12.3,由于需要保留一位小数的数字,则将变量定义为“##. #”。

日期类型用“ <yyyy/mm/dd> ”定义,表示年/月/日的意思。

注意:如果是一道选择题,则在编写 QES 文件时,还会出现第 4 部分信息,即选项解释,选项解释主要用于解释对选项的赋值情况。

对于各变量的设定,需要点击快捷菜单栏中的字段编辑器 ,如文本变量的设定,按字段编辑器,弹出如下对话框:



设定长度为 6,即三个中文字符。其他变量依次进行设定,形成如下文件:

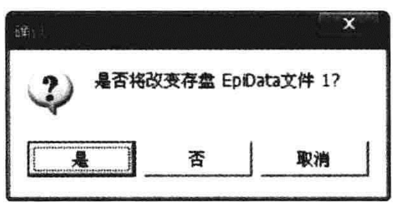
n1 姓名______
n2 年龄 # #岁
n3 性别 # 1=男 2=女
n4 是否糖尿病 # 1=是 2=不是
n5 降糖方案 # 1=只服降糖药 2=只注射胰岛素 3=降糖药+胰岛素
n6 是否高血压 # 1=是 2=不是

Step 2:**生成 REC 文件**

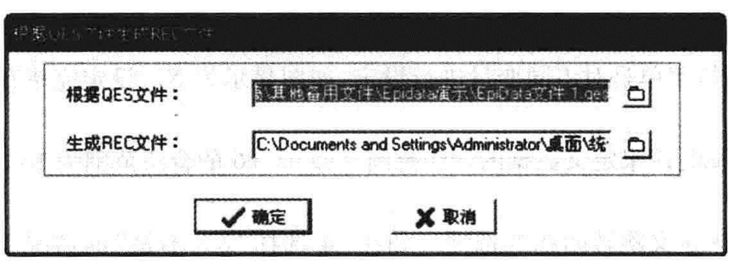
成功编写完 QES 文件后，可单击“2 生成 REC 文件”，也可以在下拉菜单中选择“数据格式预览”，确认数据的内容无误，格式美观后，再点击“2 生成 REC 文件”。



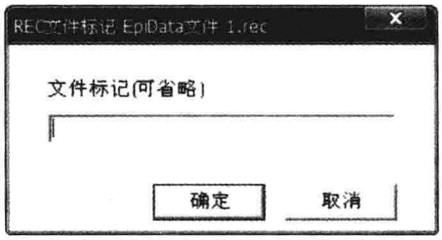
软件会自动弹出一个对话框，提示先保存 QES 文件，点击按钮“是”保存文件。



然后软件弹出对话框，可以自己根据需要设定存放 REC 文件的位置，名称默认与 QES 文件一致，单击“确定”。



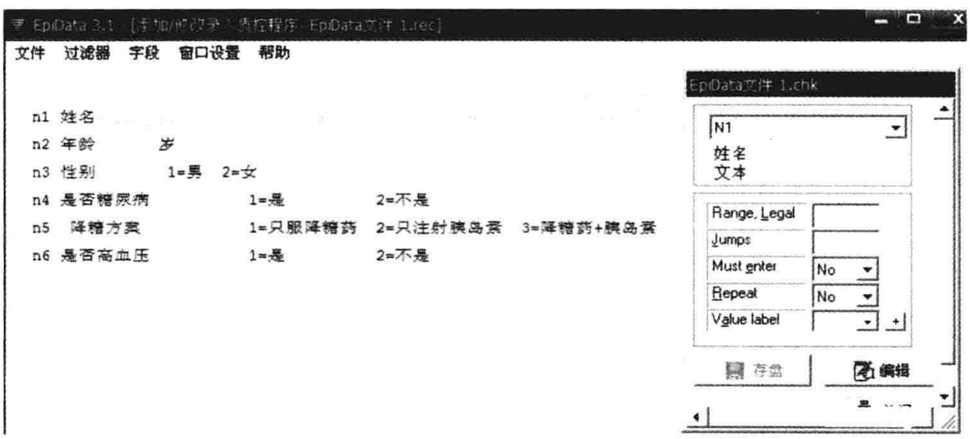
又弹出一个文件标记的对话框，可根据实际情况酌情标记文件，也可以不标记，点“确定”后就会弹出 REC 文件已生成的信息，点击“确定”，就成功建立了关于该调查表的 REC 文件。





Step 3:**建立 CHK 文件**

单击“3 建立 CHK 文件”,打开刚才生成的 REC 文件,则出现如下图所示界面。CHK 文件是录入质控文件,用于在录入数据的过程中进行质量控制,以保证数据录入的合法性和准确性。能够对输入的数据进行逻辑查错是 EpiData 作为专业数据录入软件的强大之处。



下面就界面右边对话框中的主要常用选项进行详细说明:

1. 第一个下拉菜单是对不同变量进行设定,如图显示为 N1,设定完毕存盘后再选 N2 进行设定。

2. Range, Legal:用来定义数据的合法范围。如 n3,n6 的合法范围为 1 - 2;n5 的合法范围为 1 - 3。

3. Jumps:用来定义跳转问题的选项。如在 n4 选择“2 = 不是”的情况下,则应自动跳过 n5,跳转至 n6。因此,在 n4 的 Jump 项里输入“2 > n6”表示选择 2 就跳到 n6。

4. Must enter:有“YES”和“NO”两个选项,选择“YES”则表示该题目必须录入,否则就有可能会出现缺失值。

5. Repeat:指重复出现上一条记录的值。例如,如果上一份问卷录入时 n3 选择的是“2 =

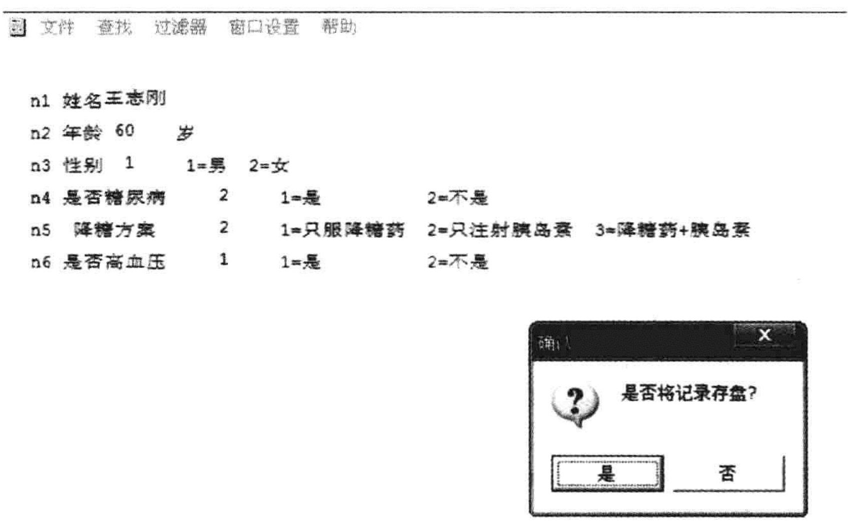
男性”,那么在录入下一份问卷时,系统会自动在 n3 问题的黄色框中出现 2,重复上一条记录的值。

6. Value label:对于分类数据定义其值的含义如 1 =“男”,2 =“女”。

将各个变量设置好后,点击“存盘”,然后“关闭”,这就建立了 CHK 文件,可在下面的数据录入阶段对整个过程和录入的数据进行质量控制。

Step 4:**数据录入**

左键单击“4 数据录入”,出现“打开”对话框,打开之前生成的 REC 文件,即可进行数据录入。每一条记录输完后,会自动弹出“是否将记录存盘”的对话框如下图,单击“是”,记录就可自动保存并进入到下一条记录。



数据双录入是指两个人独立对同一数据资料进行录入,然后对两个人的录入结果进行差异比对并纠错,以保证数据的准确性和可靠性。双录入的具体操作是先将建立好的数据库备份到另外一个盘或电脑,然后双人录入,数据录入完成后,在数据处理阶段,利用“一致性检验”进行差异比对,然后对照源数据,修改不同的数据即可。

Step 5:**数据处理**

点击其中某项后,弹出“打开”对话框,打开需要处理的 REC 文件即可。出现下拉菜单。

1. 显示 REC 文件结构　点击后会显示出该 REC 文件的大小、最近版本、字段数、记录数、是否使用质控程序,以及在 REC 文件中的字段(编号,名,变量标记,字段类型,宽度,录入质控)等信息。

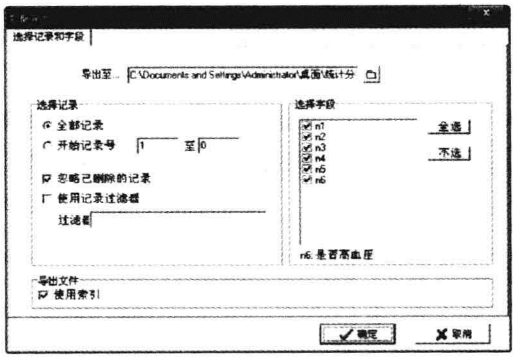
2. 数据一览表　可查看已经录入的数据情况。

3. 一致性检验　在双录入数据后,比较两次录入结果的差异情况时使用。

Step 6:**数据导出**

数据录入完成后,需将数据导出,以便采用统计软件进行分析。EpiData 提供多种软件支持,可将数据导出为 TXT、DBF、Stata、Excel、SPSS、SAS 类型的文件,也可将其导出为另一 REC 文件。数据导出后,就可用相应的软件对其进行统计分析处理了。下面介绍如何导出为 SPSS 数据库。

在“6 数据导出“的下拉菜单中,选”导出为 SPS 文件(SPSS 文件)“,系统自动跳出对话框,让你选择 REC 文件,接着对 REC 文件中导出的内容进行选择,如下图所示,一般不用选择,只是需要注意导出的位置,以利于查找。



Step 7:**在 SPSS 中生成 SPSS 数据库**

EpiData 生成 SPSS 文件并非是我们熟悉的 SPSS 数据文件(.sav 文件),而是 SPSS 语法文件(.SPS 文件),所以我们需要在 SPSS 软件中进行以下操作:

- 文件(F)—打开(O)—语法(S)
- 在弹出的对话框中,打开由 EpiData 生成的 SPS 文件
- 弹出 SPSS 语法编辑窗口,该窗口里出现如下提示:

*********** 将下一行解除注释,并存盘 *****************.

* SAVE OUTFILE="C:\Documents and Settings\Administrator\桌面\统计分析书稿\其他备用文件\Epidata 演示\演示文件.sav".

**.

- 按照提示,去除注释,如下所示:

SAVE OUTFILE="C:\Documents and Settings\Administrator\桌面\统计分析书稿\其他备用文件\Epidata 演示\演示文件.sav".

- 窗口内容全部选中,按运行按钮▶,则在数据窗口生成数据库

注意事项:7 个步骤并非都是必需的,Step 3 建立 CHK 文件为质量控制设置,Step 5 数据处理为双录入检测,因此可以省略。

中篇

统计分析篇（SPSS 18 中文版）

>> 第四章 t 检验

《隋唐演义》里的程咬金只有三板斧，而我们很多科研工作者面对数据却只有两板斧：t 检验和卡方检验，这说明我们的统计知识普及还任重道远，同时也说明了这两种方法的重要性。本章闪亮登场的就是大名鼎鼎的 t 检验。

t 检验是最简单的假设检验(hypothesis testing)方法，也是统计学中跨里程碑式的一个杰作。最早是由 W. S. Gosset 在 1908 年以笔名"Student"发表的一篇关于 t 分布的论文中提出，开创了小样本计量资料进行统计推断的先河。

一、t 检验的适用条件

1. 随机样本，即数据的独立性。
2. 来自正态分布的总体，即各样本所代表的总体呈正态分布。
3. 各样本所代表的总体方差齐。

二、SPSS 主要在**分析**下拉菜单中的**比较均值**项进行 t 检验

所包含的具体统计过程如下：

1. 均值　该过程实际上更倾向对样本进行描述，它可以对需要比较的各组进行统计描述，进行检验前的预分析。
2. 单样本 t 检验　进行样本均数与已知总体均数的比较。
3. 独立样本 t 检验　进行两样本均数差别的比较，即通常所说的两组资料 t 检验。
4. 配对样本 t 检验　进行配对资料的均数比较，即配对 t 检验。
5. 单因素 ANOVA　进行多组样本均数的比较，即成组设计的方差分析。

本章只介绍单样本 t 检验、配对样本 t 检验和独立样本 t 检验，单因素 ANOVA 在方差分析章节中进行讲解。

第一节　单样本 t 检验

一、方法原理

单样本 t 检验研究样本均数与总体均数是否相等，可以构成如下假设检验：

$H_0: \mu = \mu_0$，样本均数与假定总体均数的差异完全由抽样误差造成。

$H_1: \mu \neq \mu_0$，样本均数与假定总体均数的差异除了由抽样误差造成外，确实也反映了实际

的总体均数与假定的总体均数间的差异。

如果 H_0成立，即样本均数与总体均数的差别仅反映抽样误差，则这种差别一般不会太大，即 $|t|$ 值不会太大；如 $|t|$ 值很大，超过了事先规定的界值，则有理由怀疑 H_0的成立。

二、分析示例

10 例男性矽肺患者的血红蛋白(g/dl)如下表，已知男性健康成人的血红蛋白正常值为14.02(g/dl)，问矽肺患者的血红蛋白是否不同于一般健康成人？

病例号	1	2	3	4	5	6	7	8	9	10
血红蛋白	11.3	15.0	15.0	13.5	12.8	10.0	11.0	12.0	13.0	14.0

三、研究假设

H_0：所有矽肺患者的血红蛋白均数 = 14.02。

H_1：所有矽肺患者的血红蛋白均数 ≠ 14.02。

a = 0.05

四、数据录入

(一)变量视图

名称 x 标签 血红蛋白

	名称	类型	宽度	小数	标签	值
1	x	数值(N)	8	1	血红蛋白	无

(二)数据视图

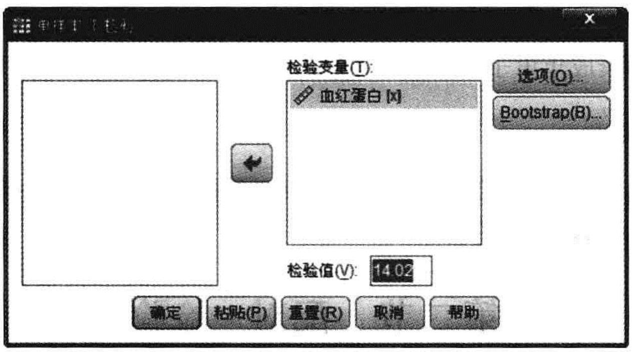
	x
1	11.3
2	15.0
3	15.0
4	13.5
5	12.8
6	10.0
7	11.0
8	12.0
9	13.0
10	14.0

五、操作流程

分析—比较均值—单样本 t 检验
检测变量(T):血红蛋白 x
检验值(V):键入 14.02

确定

这是单样本 t 检验的主对话框,非常简单,将变量血红蛋白 x 选入检验变量(T),检验值(V)中填入需要比较的总体均数,此处为 14.02,其他均按默认值即可。



六、结果解释

1. 正态分布的连续变量一般采用均数和标准差相结合的方式进行描述,本例当中为 12.763 ± 1.6892。

单个样本统计量

	N	均值	标准差	均值的标准误
血红蛋白	10	12.763	1.6892	0.5342

此处注意标准差和标准误的区别与联系。

(1)区别

①含义不同:标准差描述了观察值之间的变异度大小,标准差越大,则观察值越分散;标准差越小,则观察值越集中,均数的代表性就越强。标准误描述了样本均数之间的变异度大小,标准误越大,样本均数与总体均数差异越大,抽样误差越大。

②用途不同:标准差表示观察值的变异度大小,标准误反映抽样误差的大小。

(2)联系:两者均是变异度指标,样本均数的标准差即标准误,两者呈正比。

注意:我们在进行统计描述时,一般采用均数和标准差相结合的形式,但是由于标准差数值上大于标准误,如本例标准差为 1.6892,而标准误则为 0.5342,所以有人偷梁换柱,用均数

和标准误相结合的形式描述资料，这种做法欠妥。

2. 下表为单样本 t 检验表，从左到右依次为 t 值(t)、自由度(df)、P 值(Sig 双侧)、两均数的差值(Mean Difference)、差值的 95% 可信区间(95% Confidence Interval of the Difference)。

单个样本检验

	检验值 = 14.02					
	t	df	Sig.(双侧)	均值差值	差分的 95% 置信区间	
					下限	上限
血红蛋白	-2.353	9	.043	-1.2569	-2.465	-.048

3. 综合两表，统计描述 $\bar{x} \pm s = 12.8 \pm 1.69$，统计推断 $t = -2.353$，$P < 0.01$，因此拒绝 H_0，接受 H_1，认为矽肺患者与健康成年男子的血红蛋白不同，结合具体的均数值，可以认为矽肺患者的血红蛋白低于健康成年男子。

七、注意事项

单样本 t 检验是一种非常稳健的统计方法，只要数据分布不呈强烈偏态分布，也没有明显的极端值，其分析结果都是稳定的。

第二节 配对 t 检验

一、方法原理

配对 t 检验是单样本 t 检验的扩展，其原理为将配对设计的差值均数与总体均数 0 进行比较，常见的配对设计见于以下 4 种情况：①同一受试对象处理前后的数据；②同一受试对象两个不同部位的数据；③同一样品用两种方法(仪器等)检验的结果；④配对的两个受试对象分别接受两种处理。其目的在于推断两种处理(方法等)的结果有无差别。

二、分析示例

某地区随机抽取 12 名贫血儿童家庭，实行健康教育干预三个月，干预前后儿童的血红蛋白(%)测量结果如表所示，试问干预前后该地区贫血儿童血红蛋白(%)平均水平有无变化？(《卫生统计学》第六版例 7-2，P104)。

干预前血红蛋白	36	46	53	57	65	60	42	45	25	55	51	59
干预后血红蛋白	45	64	66	57	70	55	70	45	50	80	60	60

三、研究假设

H_0：干预前后血红蛋白差值的总体均数为 0，即该干预对血红蛋白值的改变无影响。

H_1:干预前后血红蛋白差值的总体均数不为0,即该干预对血红蛋白值的改变有影响。

a =0.05

四、数据录入

(一)变量视图

名称 id 标签 序号

名称 x1 标签 干预前

名称 x2 标签 干预后

	名称	类型	宽度	小数	标签	值
1	id	数值(N)	8	0	序号	无
2	x1	数值(N)	8	0	干预前	无
3	x2	数值(N)	8	0	干预后	无

(二)数据视图

	id	x1	x2
1	1	36	45
2	2	46	64
3	3	53	66
4	4	57	57
5	5	65	70
6	6	60	55
7	7	42	70
8	8	45	45
9	9	25	50
10	10	55	80
11	11	51	60
12	12	59	60

五、操作流程

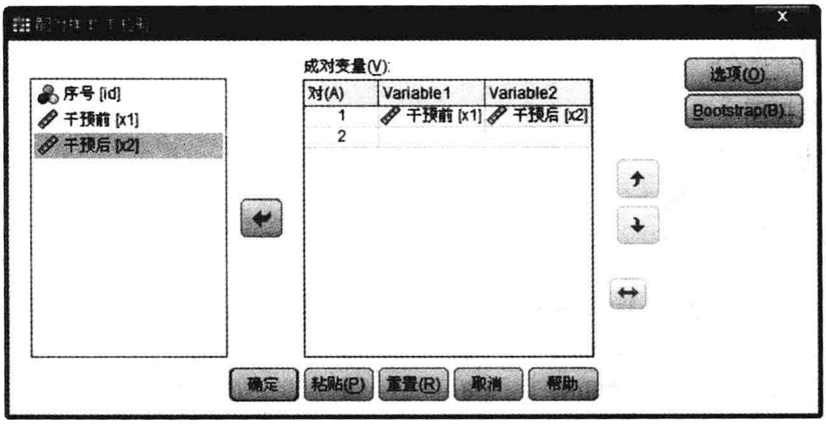
分析—比较均值—配对样本 *t* 检验

成对变量(V): 干预前 x1 干预后 x2

确定

下图为配对样本 *t* 检验的主对话框,虽然配对样本 *t* 检验的过程实际上和单样本 *t* 检验过程重复(等价于已知总体均数为0的情况),但配对样本 *t* 检验过程采用的数据输入格式和前者不同,因此仍有存在价值。主对话框非常简单,只需要将配对的两个变量选入即可(按计算

机 CTRL 键同时选上 x1 和 x2 两个变量即可）。



六、结果解释

1. 干预前后两组的统计描述，干预前 $\bar{x}\pm s=49.5\pm11.33$ 干预后 $\bar{x}\pm s=60.2\pm10.60$。

成对样本统计量

		均值	N	标准差	均值的标准误
对 1	干预前	49.50	12	11.334	3.272
	干预后	60.17	12	10.599	3.060

2. 此处给出了配对变量间的相关性分析，即 Pearson 相关系数，但此处意义并不大。

成对样本相关系数

		N	相关系数	Sig.
对 1	干预前 & 干预后	12	.482	.112

3. 下表是对 x1（干预前）－x2（干预后）形成的新的变量进行单样本 t 检验，其假设检验为差值是否为 0，其差值的均数为 －10.667，标准差为 11.179，差值的 95% 的可信区间为（－17.769，－3.564），$t=-3.305$，$P=0.007<0.01$，故可以认为健康教育前后该地区儿童血红蛋白（%）有变化，且血红蛋白（%）有所增加。

成对样本检验

		成对差分					t	df	Sig.（双侧）
		均值	标准差	均值的标准误	差分的 95% 置信区间				
					下限	上限			
对 1	干预前 －干预后	－10.667	11.179	3.227	－17.769	－3.564	－3.305	11	.007

七、注意事项

本设计属于一个有缺陷的科研设计，因为治疗前后的血红蛋白改变还受到时间因素的影响，重新设计研究方案时可以增加设置同期对照，如安慰剂对照来排除时间因素的影响，但统计方法也随之发生改变。

第三节　成组 t 检验

一、方法原理

有时无法将受试对象逐个配对，可将受试对象随机分成两组，每组接受不同处理，检验两组均数，以达到比较的目的。成组 t 检验可以构成如下假设检验：

$H_0:\mu_1=\mu_2$，两个样本均数的差异完全是由抽样误差造成的，两个总体均数相同。

$H_1:\mu_1\neq\mu_2$，两个样本均数的差异除由抽样误差造成外，两个总体均数确实存在差异。

由于 H_0 假设的是两样本来自同一总体，因此两样本 t 检验在推导过程中除了要求总体服从正态分布外，还要求两样本总体的方差相同，如果这些应用条件不满足，情况较轻时可以采用校正 t 检验的结果，否则应使用变量转换使之满足条件，或采用非参数检验过程。

二、分析示例

某妇产医院的研究者欲探讨孕妇在孕期补充钙制剂对血清骨钙素(ng/ml)的影响，选取孕妇的年龄，基础骨钙素值接近，孕周在 26 ~ 28 周的 30 名孕妇，随机分成两组，每组 15 人。试验组孕妇补充选定的某种钙制剂，对照组孕妇采用传统膳食，产后 40 ~ 50 天内测定两组孕妇血清骨钙素的改变值，结果如下(《卫生统计学》第六版例 7 - 4，P106)：

试验组	10.2	8.9	10.1	9.2	-0.8	10.6	6.5	11.2	9.3	8.0	10.7	9.5	12.7	14.4	11.9
对照组	5.0	6.7	-1.4	4.0	7.1	-0.6	2.8	4.3	3.7	5.8	4.6	6.0	4.1	5.1	4.7

三、研究假设

H_0：两组产妇骨钙素改变值的总体均数相等

H_1：两组产妇骨钙素改变值的总体均数不等

a = 0.05

四、数据录入

(一)变量视图

<u>名称</u> group <u>标签</u> 组别

<u>名称</u> x　　<u>标签</u> 骨钙素改变值

	名称	类型	宽度	小数	标签	值
1	group	数值(N)	8	0	组别	无
2	x	数值(N)	8	1	骨钙素改变值	无

（二）数据视图（部分）

	group	x
1	1	10.2
2	1	8.9
3	1	10.1
4	1	9.2
5	1	-.8
6	1	10.6
7	1	6.5
8	1	11.2
9	1	9.3
10	1	8.0
11	1	10.7
12	1	9.5
13	1	12.7

五、操作流程

分析—比较均值—独立样本 t 检验

检验变量(T)　：骨钙素改变值 x

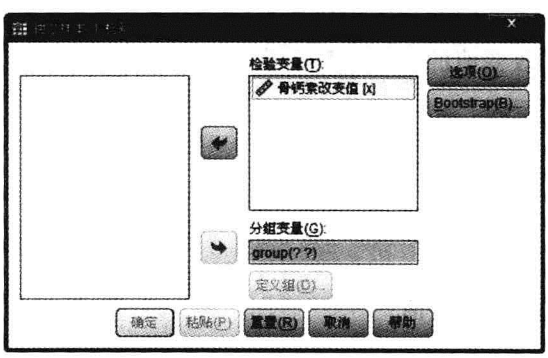
分组变量（G）：组别 group

选中变量 group: 定义组（D）

使用指定值（U）组 1：键入 1；　组 2: 键入 2 —— 继续

确定

1. 这是独立样本 t 检验的主对话框，也非常简单明了，检验变量(T)为骨钙素改变值 x，分组变量(G)为 group，不过分组变量需要按定义组(D)进行进一步定义，见下图。



2. 这是定义组的对话框,本例当中明确了两组的具体赋值1和2,只需要填入即可。如果分组变量赋值不明确,而需要按照某个取值的分界线来进行比较,可以采用切割点(cut point)来指定,在此指定分界值,系统将会将记录自动分成小于分界值和大于等于分界值的两组来进行比较。



六、结果解释

1. 下表给出了两组的统计量描述值,试验组(group = 1) $\bar{x} \pm s = 9.5 \pm 3.43$,对照组(group = 2) $\bar{x} \pm s = 4.1 \pm 2.37$。

组统计量

	组别	N	均值	标准差	均值的标准误
骨钙素改变值	1	15	9.493	3.4250	.8843
	2	15	4.127	2.3741	.6130

2. 下表分为两大部分,第一部分为Levene方差齐性检验,用于判断两总体方差是否齐,这里的检验结果为 $F = 0.374$, $P = 0.546 > 0.05$,可见在本例中方差齐;第二部分则给出了两组在总体方差齐和不齐两种情况下的 t 检验结果,由于前面的方差齐性检验结果为方差齐,第二部分就选用方差相等时的 t 检验结果,即上面一行列出的 $t = 4.988$, $P < 0.01$(注意,此处表格虽然标注 P 值为0.000,但并不意味着 $P = 0$,而是由于受到显示的小数位限制,所以必须标注 $P < 0.01$)。从而认为两组产妇骨钙素改变值的总体均数有差异,且孕期补钙者骨钙素改变值较大。

独立样本检验

		方差方程的 Levene 检验		均值方程的 t 检验						
		F	Sig.	t	df	Sig.（双侧）	均值差值	标准误差值	差分的 95% 置信区间 下限	差分的 95% 置信区间 上限
骨钙素改变值	假设方差相等	.374	.546	4.988	28	.000	5.3667	1.0760	3.1626	7.5708
	假设方差不相等			4.988	24.930	.000	5.3667	1.0760	3.1503	7.5831

七、注意事项

成组 t 检验要求数据来源于方差相同的正态总体，正态性要求分组考虑，而非合并考察，t 检验对于资料的正态性有一定的耐受能力，如果资料只是少许偏离正态，结果仍然稳健，但如果数据分布严重偏离正态，均数不能很好代表数据的集中趋势，才考虑数据变换或非参数检验。

>>第五章　方差分析

一、基本思想

方差分析(analysis of variance, ANOVA)由英国著名统计学家 R. A. Fisher 提出,又称变异数分析或 F 检验,是一种以 F 值为统计量的计量资料的假设检验方法,它是将总方差(严格的说是离均差平方和)分解成两个或多个部分方差和,总自由度分解成相应各部分自由度为手段,目的在于推断两组或多组的总体均数是否相等或检验两个或多个样本均数间的差异是否有统计学意义。

二、基本概念

1. 总变异(total variation)　全部观察值或变量值大小不同,这种变异称为总变异。

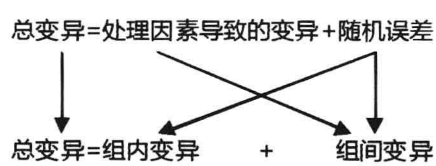
2. 组间变异(variation between groups)　指各个处理组样本均数不同,与总体均数也不同,组间变异产生的原因包括处理因素和随机误差(含个体差异和测量误差)。

3. 组内变异(variaition within groups)　各个处理组内部观察值或变量值大小不同,这种变异称为组内变异,组内变异反映了部分随机误差。

4. 处理因素导致的变异　由于研究目的的需要,我们对不同的试验组施加不同影响,如采用不同的治疗方法,从而使各组之间的均数出现差异,这是我们希望看到的变异。

5. 随机误差　由于试验对象的个体差异和测量误差等原因,使不同观察对象出现不同的测量结果,这是我们不希望看到的变异,但又是不可避免的因素。

三、基本概念之间的关系



如图所示,我们需要了解的是处理因素导致的变异和随机误差,但是我们能够测量的却是组内变异和组间变异,因此我们了解总变异的构成,我们对此计算统计 F 值:

$$F = 组间变异 \div 组内变异$$

如果处理因素导致的变异存在,则 F 值比 1 大得多。

如果处理因素导致的变异不存在,则理论上 $F=1$,即组间变异和组内变异均只包括随机

误差。但实际上，F 值不正好等于 1，而是在 1 的周围波动，但不会是负数。

四、SPSS 方差分析模块

SPSS 主要在分析下拉菜单项中的一般线性模型（General Linear Model，GLM）中实现，其包含的具体统计过程如下：

1. *单变量*（Univariate） 当因变量只有一个时，采用单变量方差分析，完全随机设计的方差分析、随机区组设计的方差分析、析因设计的方差分析均采用此项，堪称方差分析中的大哥大。

2. *多变量*（Multivariate） 当因变量不止一个时，采用多变量方差分析。

3. *重复度量*（Repeated Measures） 很多试验中对同一受试对象进行多次观察，这样的数据称为重复数据，重复数据需采用重复度量的方差分析。大家对这种数据经常误用单变量方差分析或多变量方差分析，其结果是不准确的。

4. *方差分量估计*（Variance Compoenents） 用于对层次数据拟合方差成分模型，它是普通线性模型向随机效应的进一步扩展，是一种可以考察各个层次因素的变异大小，从而为那些层次上可能存在组内聚集性、如何可能减少数据变异提供信息的统计方法，也就是现在热门的多水平模型的最原始形式。

但本章将完全随机设计资料的方差分析采用比较均值项中的单因素 ANOVA 过程进行运算，当然也可采用一般线性模型项中的单变量过程来计算。

第一节　完全随机设计资料的方差分析

一、方法原理

完全随机设计（completely randomized design）是将同质的受试对象随机地分配到各处理组，再观察其试验效应。各组样本含量可以相等，也可以不相等。完全随机设计常见的统计学检验为单因素方差分析（one-way ANOVA），所解决的是一个因素（factor）之下的多个不同水平（level）之间的关系问题，可视为独立样本 t 检验的扩展，因为独立样本 t 检验是分析一个因素下两个水平之间的关系。从 SPSS 运算模块将单因素方差分析与独立样本 t 检验一同归于比较均数模块也可以看出其关系。

二、分析示例

为研究钙离子对体重的影响作用，某研究者将 36 只肥胖模型大白鼠随机分为 3 组，每组 12 只，分别给予高脂正常剂量钙（0.5%），高脂中剂量钙（1.0%）和高脂高剂量钙（1.5%）三种不同的饲料，喂养 9 周，测其喂养前后体重的差值。问三组不同喂养方式下大白鼠体重改变是否不同？（《卫生统计学》第六版例 8-1，P123）

三种不同喂养方式下大白鼠体重喂养前后差值

正常钙 (0.5%)	中剂量钙 (1.0%)	高剂量钙 (1.5%)
332.96	253.21	232.55
297.64	235.87	217.71
312.57	269.3	216.15
295.47	258.9	220.72
284.25	254.39	219.46
307.97	200.87	247.47
292.12	227.79	280.75
244.61	237.05	196.01
261.46	216.85	208.24
286.46	238.03	198.41
322.49	238.19	240.35
282.42	243.49	219.56

三、研究假设

H_0:三组不同喂养方式下大白鼠体重改变的总体平均水平相同。

H_1:三组不同喂养方式下大白鼠体重改变的总体平均水平不全相同。

a = 0.05

四、数据录入

1. 变量视图

名称 group 标签 喂养组别

名称 x 标签 体重差值

	名称	类型	宽度	小数	标签	值
1	group	数值(N)	8	0	喂养组别	无
2	x	数值(N)	8	2	体重差值	无

2. 数据视图(部分)

	group	x
1	1	332.96
2	1	297.64
3	1	312.57
4	1	295.47
5	1	284.25
6	1	307.97
7	1	292.12
8	1	244.61
9	1	261.46
10	1	286.46
11	1	322.49
12	1	282.42
13	2	253.21

五、操作流程

分析—比较均值—单因素 ANOVA

因变量列表(E):体重差值 x

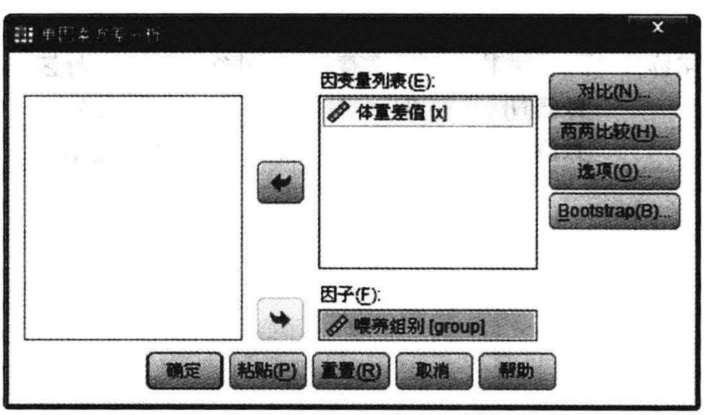
因子(F):喂养组别 group

两两比较(H):☑ LSD ☑ S-N-K—继续

选项(O):☑描述性 ☑方差同质性检验—继续

确定

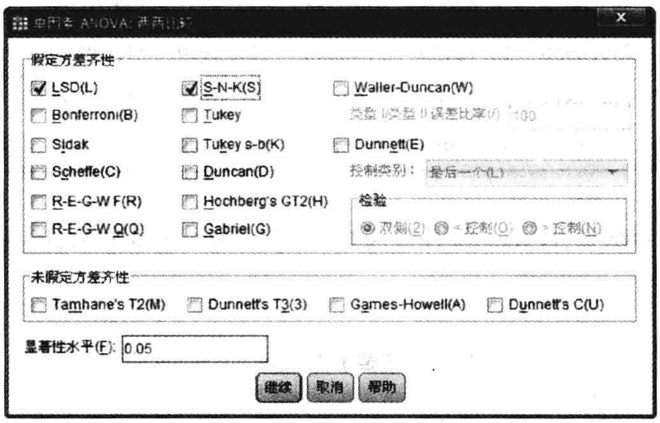
1. 下图是单因素方差分析的主对话框,其形式也与独立样本 *t* 检验类似,我们将因变量体重差值 x 选入因变量列表(E)中,再将分组变量喂养组别 group 选入因子(F)中。



2. 如果总体上有差别,即代表三种不同喂养模式下大白鼠体重改变的总体平均水平不全相等,那么到底是哪些组之间的体重改变的总体平均水平不相等,我们采用两两比较的方法得出结果。两两比较的方法很多,其选择依赖于研究设计和数据类型,有以下几点参考意见:

验证性研究　即在设计阶段就根据研究目的或专业知识而计划好的某些均数的两两比较,常用 LSD 法或 Bonferroni 法。

探索性研究　即研究阶段未计划进行多重比较,在方差分析得到有统计意义的 F 值后,进行两两比较。若进行多个试验组与一个对照组的比较时,用 Dunnett 法;进行任一两组比较且样本含量相同,用 Tukey 法或 SNK 法;若进行任一两组比较但样本含量不同,用 Scheffe 法,但此法比较保守,即可能总的差别有意义但找不出有差异的任何两个样本。



3. 下图"选项"对话框中描述性(D)主要给出几组数据的均数和标准差,而方差同质性检验(H)即方差齐性检验。



六、结果解释

1. 各组数据的统计描述，正常钙组(group = 1)$\bar{x} \pm s = 293.36 \pm 24.62$，中剂量钙组(group = 2)$\bar{x} \pm s = 239.49 \pm 18.72$，高剂量钙组(group = 3)$\bar{x} \pm s = 224.78 \pm 23.24$。

描述

体重差值

	N	均值	标准差	标准误	均值的 95% 置信区间		极小值	极大值
					下限	上限		
1	12	293.3683	24.62068	7.10738	277.7251	309.0116	244.61	332.96
2	12	239.4950	18.72159	5.40446	227.5999	251.3901	200.87	269.30
3	12	224.7817	23.24461	6.71014	210.0127	239.5506	196.01	280.75
总数	36	252.5483	36.93959	6.15660	240.0498	265.0469	196.01	332.96

2. 方差齐性检验结果，Levene 统计量为 0.319，$P = 0.729 > 0.05$，因此可以认为各样本所在总体的方差齐。

方差齐性检验

体重差值

Levene	统计量	df1	df2	显著性
	.319	2	33	.729

3. 这是最重要的一个统计结果，给出了单因素方差分析的结果，即常说的各组总体上是否有差异，本表中 $F = 31.355$，$P < 0.01$，因此可以认为三种不同喂养方式下大白鼠体重改变不全相同。特别注意其字眼“不全相同”，即各组当中只要有两组出现差异，则可认为总体有差异。

ANOVA

体重差值

	平方和	df	均方	F	显著性
组间	31291.796	2	15645.898	31.355	.000
组内	16466.867	33	498.996		
总数	47758.663	35			

4. 如果总体有差异，那么大家关心各组之间的情况，需要进行两两比较，也称为“事后比较(Post Hoc)”，下面给出了两种常见的两两比较的方法：LSD 法和 SNK 法，大家在实际中选用其中一种即可。

多重比较(LSD 法)

因变量:体重差值

	(I) 喂养组别	(J) 喂养组别	均值差 (I-J)	标准误	显著性	95% 置信区间	
						下限	上限
LSD	1	2	53.87333*	9.11954	.000	35.3195	72.4272
		3	68.58667*	9.11954	.000	50.0328	87.1405
	2	1	-53.87333*	9.11954	.000	-72.4272	-35.3195
		3	14.71333	9.11954	.116	-3.8405	33.2672
	3	1	-68.58667*	9.11954	.000	-87.1405	-50.0328
		2	-14.71333	9.11954	.116	-33.2672	3.8405

* 均值差的显著性水平为 0.05

LSD 法(最小二乘法)的两两比较结果比较清楚,3 组两两比较有 3 组数据:
正常钙组(group =1)和中剂量钙组(group =2)比较:均数差为 53.87,$P<0.01$
正常钙组(group =1)和高剂量钙组(group =3)比较:均数差为 68.59, $P<0.01$
中剂量钙组(group =2)和高剂量钙组(group =3)比较:均数差为 14.71,$P=0.116>0.05$

体重差值(SNK 法)

	喂养组别	N	a =0.05 的子集	
			1	2
Student-Newman-Keuls[a]	3	12	224.7817	
	2	12	239.4950	
	1	12		293.3683
	显著性		.116	1.000

将显示同类子集中的组均值

a. 将使用调和均值样本大小 =12.000

这是 SNK 法的两两比较结果,这个表格比较令人费解,简单说来,该表在纵向将各组均数按从小到大进行排列,高剂量组(group =3)均数 224.7817 最小,居于顶端;正常剂量组(group =1)均数为 293.3683 最大,居于底端。而横向则被分成若干亚组,不同亚组间的 P 值 <0.05,而同一亚组间的各组均数比较的 $P>0.05$,可见 SNK 法无法获知准确 P 值,只能知道两组比较的 P 值是否 >0.05。本例当中亚组 1 包括了高剂量组(group =3)和中剂量组(group =2),因此两者比较的 $P>0.05$;而正常剂量组(group =1)位于亚组 2,亚组 1 和亚组 2 之间的 $P<0.05$,因此正常剂量组(group =1)和中剂量组(group =2)比较 $P<0.05$,正常剂量组(group =1)和高剂量组(group =3)比较 $P<0.05$。

七、注意事项

着重谈谈方差分析的适用条件：

1. 独立性　准确表述为研究对象是来自所研究因素各水平之下的独立随机抽样（independence），这个条件要求最为严格，因为违反了独立性，则方差分析的误差项的固定也随之发生改变，不能简单以组内变异作为误差项。独立性的考察一般可以分析研究设计及其数据收集方法来解决，最常见的违反独立性的设计为对同一观察对象多个时间点进行观察，此种设计可采用重复测量的方差分析。

2. 正态性　准确表述为每个水平下的因变量应当服从正态分布（normality），需要注意的是正态性和方差齐性的检验是针对所有单元格（单元格指的是各研究因素各水平的交叉组合）而言，并非就整体而言，单因素方差分析就是指各组的数据都符合正态分布。Box 和 Anderson 等人的研究表明，在正态性得不到满足时，方差分析的结论并不会受到太大影响，也就是说，方差分析对于正态性的要求是稳健的（robust）。

3. 方差齐性　方差齐性也是针对所有单元格而言的，在各组间样本含量相差不太大时，方差轻微不齐仅会对方差分析的结论有少许影响，一般而言，只要最大/最小方差之比小于 3，分析结果都是稳定的。

在方差分析中，各组在样本含量上的均衡性将会为分析计算提供很大的便利，也在一定程度上弥补了正态性或方差齐性得不到满足时对检验效能所产生的影响，这一点在多因素分析时体现得尤为明显。因此，在试验设计时应当注意均衡性的问题，当然，很多均衡性的试验设计在执行过程中产生了不均衡的数据。

第二节　随机区组设计资料的方差分析

一、方法原理

随机区组设计（randomized block design）又称配伍组设计，它通常将受试对象按性质（如动物的性别、体重、病人的病情、性别、年龄等非实验因素）相同或相近组成 b 个区组，每个区组中的 k 个受试对象分别随机分配到 k 个处理组中去。区组选择的原则在于区组间差别越大越好，区组内差别越小越好。随机区组设计将受试对象分组后，进一步控制了个体差异，因此其检验效能高于完全随机设计的方差分析。

二、分析示例

为探索丹参对肢体缺血再灌注损伤的影响，将 30 只纯种新西兰实验用大白兔，按窝别相同分为 10 个区组，每个区组的 3 只大白兔随机接受三种不同的处理，即在松止血带前分别给予丹参 2 ml/kg、丹参 1 ml/kg、生理盐水 2 ml/kg，并分别测定松止血带前及松后 1 小时血中白蛋白含量（g/L），算出白蛋白的减少量如表所示，问三种处理的效果是否不同？（《卫生统计学》第六版例 8 - 2，P127）

三种方案处理后大白兔血中白蛋白减少量(g/L)

区组	丹参(2 ml/kg)	丹参(1 ml/kg)	生理盐水
1	2.21	2.91	4.25
2	2.32	2.64	4.56
3	3.15	3.67	4.33
4	1.86	3.29	3.89
5	2.56	2.45	3.78
6	1.98	2.74	4.62
7	2.37	3.15	4.71
8	2.88	3.44	3.56
9	3.05	2.61	3.77
10	3.42	2.86	4.23

三、研究假设

研究问题1:三组大白兔的白蛋白减少量是否相同?

研究问题2:如果总体有差别,那么是哪些组之间出现差别?

四、数据录入

1. 变量视图

名称 b 标签 区组

名称 k 标签 处理

名称 x 标签 白蛋白减少量

	名称	类型	宽度	小数	标签	值
1	b	数值(N)	8	0	区组	无
2	k	数值(N)	8	0	处理	无
3	x	数值(N)	8	2	白蛋白减少量	无

2. 数据视图(部分)

	b	k	x
1	1	1	2.21
2	2	1	2.32
3	3	1	3.15
4	4	1	1.86
5	5	1	2.56
6	6	1	1.98
7	7	1	2.37
8	8	1	2.88
9	9	1	3.05
10	10	1	3.42
11	1	2	2.91
12	2	2	2.64

五、操作流程

分析—一般线性模型—单变量
因变量（D）：白蛋白减少量 x
固定因子（F）：处理 k
随机因子（A）：区组 b

模型（M）：设定（C）

类型（P）：主效应

模型（M）框：b k — 继续

两两比较（H）：两两比较检验（P）：k

☑ SNK—继续

选项（O）：☑描述统计（D）—继续

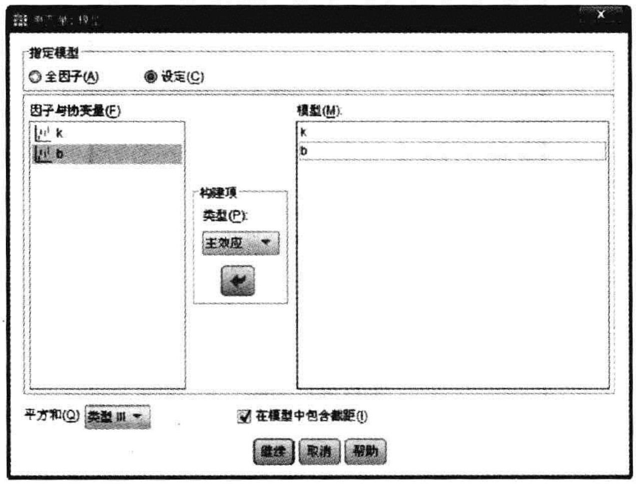
确定

1. **因变量(D)**：选入需要分析的因变量(dependent variable)，因变量通常指我们所关心的测量变量，如本例当中的白蛋白减少量 x。

固定因子(F)和随机因子(A)：两者均指我们的处理因素和区组因素，也就是通常所说的自变量。固定因子(fixed factor)指该因子在样本中所有可能的水平都出现了，即该因子的所有可能水平均列出了，无需进行外推，如本例三种方案处理大白兔，有且只有三种方案，所以属于固定因子。而随机因子(random factor)指该因子的所有可能的水平在样本中没有都出现，需要进行外推，如本例中按窝别分为 10 个区组，我们认为窝别不止 10 种，因此属于随机因子。可见固定因子和随机因子是由试验设计决定的，所以我们可以根据试验设计的不同，将同一个因素视为固定因子或随机因子均可。



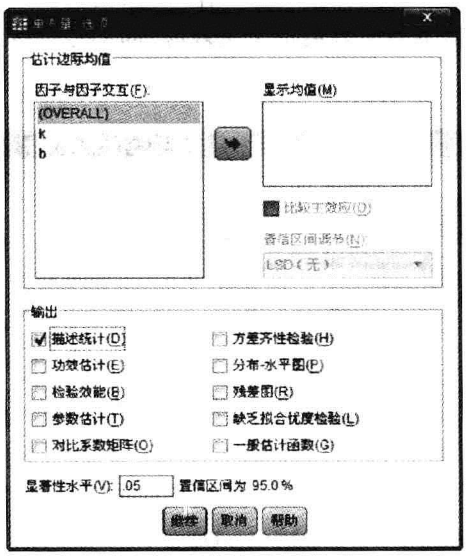
2. 下图为方差分析模型设定部分,又可以分为全模型(full model)和设定模型(specified model),全模型包括所有自变量的主效应和交互作用,而设定模型则自由设定。在随机区组设计当中需要自己设定模型,模型当中只包括了主效应,即不考虑区组因素和处理因素的交互作用。



3. 下图为两两比较的选项,我们只需要对处理因素之间是否存在差异感兴趣,因此在两两比较中选入 k,选用 S-N-K 法作为两两比较的统计方法。



4. 下图表示我们需要获得各种处理的均数和标准差，因此选择描述统计(D)。



六、结果解释

1. 下表给出了处理因素和区组因素的水平和例数，如处理组的水平为1(丹参2 ml/kg)、2(丹参1 ml/kg)、3(生理盐水)，各水平的例数均为10；而区组的水平从1到10，各水平的例数均为3。

主体间因子

		N
处理	1	10
	2	10
	3	10
区组	1	3
	2	3
	3	3
	4	3
	5	3
	6	3
	7	3
	8	3
	9	3
	10	3

2. 下表给出了各种组合的均数和标准差,我们所关注的是处理因素各水平的均数和标准差,其具体数值为丹参 2 ml/kg(k = 1) $\bar{x} \pm s = 2.6 \pm 0.52$,丹参 1 ml/kg(k = 2) $\bar{x} \pm s = 3.0 \pm 0.40$,生理盐水 $\bar{x} \pm s = 4.2 \pm 0.40$。

描述性统计量

因变量:白蛋白减少量

处理	区组	均值	标准 偏差	N
1	1	2.2100	.	1
	2	2.3200	.	1
	3	3.1500	.	1
	4	1.8600	.	1
	5	2.5600	.	1
	6	1.9800	.	1
	7	2.3700	.	1
	8	2.8800	.	1
	9	3.0500	.	1
	10	3.4200	.	1
	总计	2.5800	.52375	10
2	1	2.9100	.	1
	2	2.6400	.	1
	3	3.6700	.	1
	4	3.2900	.	1
	5	2.4500	.	1
	6	2.7400	.	1
	7	3.1500	.	1
	8	3.4400	.	1
	9	2.6100	.	1
	10	2.8600	.	1
	总计	2.9760	.39761	10
3	1	4.2500	.	1
	2	4.5600	.	1
	3	4.3300	.	1
	4	3.8900	.	1

续表

处理	区组	均值	标准 偏差	N
	5	3.7800	.	1
	6	4.6200	.	1
	7	4.7100	.	1
	8	3.5600	.	1
	9	3.7700	.	1
	10	4.2300	.	1
	总计	4.1700	.40061	10
总计	1	3.1233	1.03660	3
	2	3.1733	1.21150	3
	3	3.7167	.59138	3
	4	3.0133	1.04290	3
	5	2.9300	.73817	3
	6	3.1133	1.35902	3
	7	3.4100	1.19147	3
	8	3.2933	.36295	3
	9	3.1433	.58561	3
	10	3.5033	.68879	3
	总计	3.2420	.81023	30

3. 下表为一个典型的方差分析结果表，处理因素（k）方差分析结果 $F=32.636$，$P<0.01$，可认为三种不同的处理效果不同，即三个总体均数中至少有两个不同。至于三个总体均数中哪些不同，则需要用多个均数间的两两比较方法。区组因素（b）中 $F=0.824$，$P=0.602>0.05$，即不能认为10个区组的总体均数不同。

主体间效应的检验

因变量：白蛋白减少量

源		Ⅲ型平方和	df	均方	F	Sig.
截距	假设	315.317	1	315.317	1822.001	.000
	误差	1.558	9	.173[a]		
k	假设	13.702	2	6.851	32.636	.000
	误差	3.778	18	.210[b]		
b	假设	1.558	9	.173	.824	.602
	误差	3.778	18	.210[b]		

a. MS(b)

b. MS(错误)

4. 下表对主体间效应的检验作进一步解释,但是对我们来说,意义并不太大。

期望均方[a,b]

源	方差成分		
	Var(b)	Var(误差)	二次项
截距	3.000	1.000	截距, k
k	.000	1.000	k
b	3.000	1.000	
误差	.000	1.000	

a. 对于每个源,期望的平方均值等于单元格中系数之和乘以方差成分,加上一个在二次项单元中包含效应的二次项

b. 期望均方基于Ⅲ型平方和

5. 又出现了令人头疼的 SNK 法两两比较结果,下表在纵向将各组均数按从小到大进行排列,丹参 2 ml/kg(k = 1)均数 2.5800 最小,居于顶端;而生理盐水组(k = 3)均数为 4.1700 最大,居于底端。而横向则被分成若干亚组,不同亚组间的 $P<0.05$,而同一亚组间的各组均数 $P>0.05$,可见 SNK 法无法获知准确的 P 值,只能知道两组比较的 P 值是否 >0.05。本例当中亚组 1 包括了丹参 2 ml/kg(k = 1)和丹参 1 ml/kg(k = 2),因此两组比较的 $P>0.05$,而生理盐水组(k = 3)位于亚组 2,亚组 1 和亚组 2 之间的 $P>0.05$,因此丹参 2 ml/kg(k = 1)和生理盐水组(k = 3)比较 $P<0.05$,丹参 1 ml/kg 组(k = 2)和生理盐水组(k = 3)比较 $P<0.05$。

白蛋白减少量

Student-Newman-Keuls[a,b]

处理	N	子集	
		1	2
1	10	2.5800	
2	10	2.9760	
3	10		4.1700
Sig.		.069	1.000

已显示同类子集中的组均值

基于观测到的均值

误差项为均值方(错误) = .210

a. 使用调和均值样本大小 = 10.000

b. Alpha = .05

七、注意事项

1. 随机区组设计中并没有考察其正态性与方差齐性,是因为无法考察。

2. 不管区组变量是否具有统计学意义,都应当保留在方程当中。

3. 作为区组变量，应当从设计上考虑与实验因素不存在交互作用，如果不能肯定是否存在交互作用，则应当采用析因设计、正交设计等更加复杂的统计模型。

第三节　析因设计资料的方差分析

一、方法原理

析因设计（factorial design）是将两个或多个试验因素的各水平进行全面组合，能够分析各试验因素的单独效应（simple effect）、主效应（main effect）和因素间的交互效应（interaction）。完全随机设计只考虑一个试验因素，而随机区组设计添加了一个区组因素，但是区组因素并不是感兴趣的试验因素，而且该设计假定区组因素和试验因素间不存在交互作用。析因设计则考察了多个试验因素，并且能够分析各因素间的交互作用，这是析因设计最大的优点。

二、分析示例

研究者欲研究煤焦油（因素 A）以及作用时间（因素 B）对细胞毒性的作用，煤焦油含量分别为 3 μg/ml（a_1）和 75 μg/ml（a_2）两个水平，作用时间分别为 6 小时（b_1）和 8 小时（b_2）。将统一制备的 16 盒已培养好的细胞随机分为四组，分别接受 A、B 不同组合情况下的四种处理（a_1b_1、a_1b_2、a_2b_1、a_2b_2），测得处理液吸光度的值（%），结果如下（《卫生统计学》第六版例 8-3，P129）：

四种不同处理情况下吸光度的值（%）

煤焦油（3 μg/ml）a_1		煤焦油（75 μg/ml）a_2	
时间（6 小时）	时间（8 小时）	时间（6 小时）	时间（8 小时）
b_1	b_2	b_1	b_2
0.163	0.127	0.124	0.101
0.199	0.168	0.151	0.192
0.184	0.152	0.127	0.079
0.198	0.150	0.101	0.086

三、研究假设

研究问题 1：煤焦油含量为 3 μg/ml 和 75 μg/ml 吸光度的总体均数是否相等？
研究问题 2：作用时间 6 小时和 8 小时吸光度的总体均数是否相等？
研究问题 3：不同煤焦油含量对不同作用时间的吸光度测量值有无影响？

四、数据录入

1. 变量视图

名称 a 标签　煤焦油 值　1 = 3 μg/ml; 2 = 75 μg/ml

名称 b 标签　时间　　值 1 = 6 小时;2 = 8 小时

名称 x 标签 吸光度

	名称	类型	宽度	小数	标签	值
1	a	数值(N)	8	0	煤焦油	{1, 3μg/ml}...
2	b	数值(N)	8	0	时间	{1, 6小时}...
3	x	数值(N)	8	3	吸光度	无

注意:此处对值赋予了标签,这样在结果中便将值用标签来替代,一目了然。但值赋标签只对输出产生影响,并不影响统计结果。

2. 数据视图

	a	b	x
1	1	1	.163
2	1	1	.199
3	1	1	.184
4	1	1	.198
5	1	2	.127
6	1	2	.168
7	1	2	.152
8	1	2	.150
9	2	1	.124
10	2	1	.151
11	2	1	.127
12	2	1	.101
13	2	2	.101
14	2	2	.192
15	2	2	.079
16	2	2	.086

五、操作流程

分析—一般线性模型—单变量

因变量(D):吸光度 x

固定因子(F):煤焦油 a　时间 b

[模型(M)]:☑全因子(A)—[继续]

[选项(O)]:☑描述统计(D)—[继续]

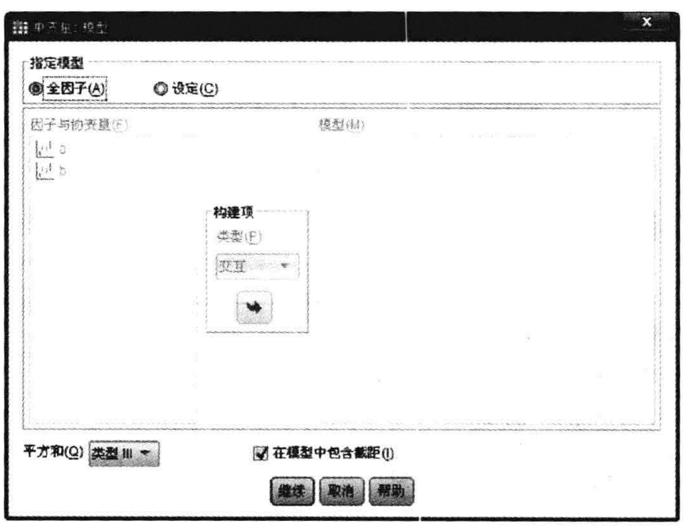
[确定]

1. **因变量(D)**:选入需要分析的因变量(dependent variable),因变量通常指我们所关心的测量变量,如本例当中的吸光度x。

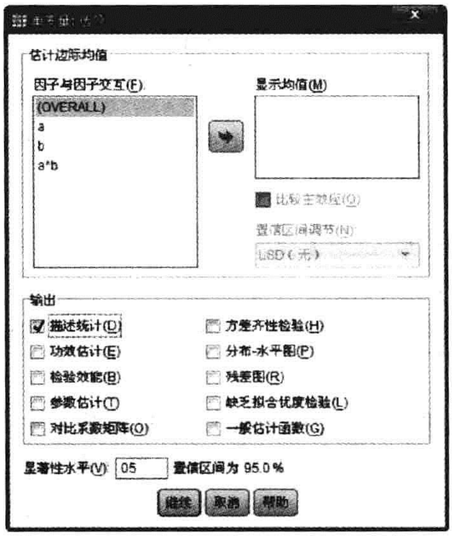
固定因子(F)和随机因子(A):即通常所说的自变量。固定因子(fixed factor)指该因子在样本中所有可能的水平都出现了,即该因子的所有可能水平均列出了,无需进行外推;而随机因子(random factor)指该因子的所有可能的水平在样本中没有都出现,需要进行外推,可见固定因子和随机因子是由试验设计决定的,所以你可以根据试验设计的不同,将同一个变量视为固定因子或随机因子均可。本例中研究者关心煤焦油含量和作用时间均无需外推,可视为固定因子。



2. 析因设计指两个或多个实验因素各水平的全面组合,即**全因子模型**,本例包括煤焦油a(主效应)、作用时间b(主效应)和两因素的交互作用。



3. 我们一般需要获知各组均数和标准差,因此选定**描述统计(D)**。



六、结果解释

1. 下表给出了两处理因素的水平和例数,如煤焦油组的水平为 1(3 μg/ml)、2(75 μg/ml),各水平的例数均为 8,而时间组的水平为 1(6 小时)、2(8 小时),各水平的例数均为 8。该表的出现是由于我们在变量视图窗口对值设定了标签,因此给出了值标签简表。

主体间因子

		值标签	N
煤焦油	1	3 μg/ml	8
	2	75 μg/ml	8
时间	1	6 小时	8
	2	8 小时	8

2. 下表给出了各种组合的描述性统计量,即 3 μg/ml + 6 小时、3 μg/ml + 8 小时、75 μg/ml + 6 小时、75 μg/ml + 8 小时、3 μg/ml、75 μg/ml、6 小时、8 小时的所有数据的均数、标准差,共计 9 组,大家根据需要进行选取。

描述性统计量

因变量:吸光度

煤焦油	时间	均值	标准 偏差	N
3 μg/ml	6 小时	.18600	.016793	4
	8 小时	.14925	.016879	4

续表

煤焦油	时间	均值	标准 偏差	N
	总计	.16763	.025077	8
75 μg/ml	6 小时	.12575	.020451	4
	8 小时	.11450	.052475	4
	总计	.12013	.037357	8
总计	6 小时	.15588	.036569	8
	8 小时	.13188	.040587	8
	总计	.14388	.039324	16

3. 最重要的方差分析结果出来了(见下表),首先看煤焦油(a)和作用时间(b)交互效应的 $F=0.696$,$P=0.421>0.05$,故尚不能认为两个因素间存在交互作用。接着看 a、b 两因素的主效应,煤焦油(a)中 $F=9.655$, $P=0.009<0.01$,认为煤焦油含量为 3 μg/ml 和 75 μg/ml 两组的总体均数不同,即煤焦油含量对吸光度有影响;而作用时间(b)中 $F=2.465$,$P=0.142>0.05$,可认为 6 小时作用时间和 8 小时两组的总体均数相同,即作用时间对吸光度无影响。

主体间效应的检验

因变量:吸光度

源	Ⅲ型平方和	df	均方	F	Sig.
校正模型	.012[a]	3	.004	4.272	.029
截距	.331	1	.331	354.335	.000
a	.009	1	.009	9.655	.009
b	.002	1	.002	2.465	.142
a * b	.001	1	.001	.696	.421
误差	.011	12	.001		
总计	.354	16			
校正的总计	.023	15			

a. $R^2=.516$(调整 $R^2=.396$)

第四节 重复测量资料的方差分析

一、方法原理

重复测量资料(repeated measurement data)是同一受试对象的同一观察指标在不同时间点上进行多次测量所得的资料,常用来分析观察指标在不同时间点上的变化。

重复测量资料与随机区组设计资料的区别:

(1)重复测量资料中同一受试对象的数据具有相关性,即不具有独立性,而随机区组设计资料同一区组内的数据不具有相关性,即具有独立性。

(2)从试验设计来看,重复测量资料中的处理因素在受试对象(看成区组)间为随机分配,但受试对象(看成区组)内的各时间点往往是固定的,不能随机分配。随机区组设计资料中每个区组内的受试对象彼此独立,处理只在区组内随机分配,同一区组内的受试对象接受的处理各不相同。

二、分析示例

某研究者欲研究青光眼结膜成纤维细胞增殖表达情况,在某医院随机抽取了 20 例青光眼患者和 24 例对照,取两组研究对象眼角膜细胞进行培养,分别在 3、7、14、21 天四个时间点观察平均细胞数(《卫生统计学》第六版例 8 - 4,P134)。

某研究者观察 20 例青光眼患者和 24 例对照的成纤维细胞数

分组	受试对象	细胞培养时间(天)			
		3	7	14	21
1	1	0.1917	1.6667	1.6500	0.7000
1	2	2.5938	2.8000	2.4000	3.8000
1	3	1.2000	4.3750	4.5250	0.2167
1	4	2.1792	1.9375	1.6688	0.6400
1	5	1.0000	1.4500	1.7750	0.4250
1	6	2.2000	4.7000	3.9667	1.7000
1	7	6.6167	5.5667	1.5625	2.5100
1	8	6.6000	4.4333	2.4125	1.6875
1	9	5.8000	4.6167	2.9875	5.9500
1	10	8.0500	7.1000	4.9500	5.0000
1	11	0.7250	2.1917	2.8000	1.7250
1	12	4.4375	6.4563	5.5250	6.9333
1	13	3.0625	7.1250	5.1000	0.9833
1	14	5.4708	4.4000	3.2500	2.2000
1	15	1.6000	0.8500	3.3750	1.8750
1	16	6.3500	8.8667	4.5000	5.9000
1	17	6.5500	6.5667	3.8375	5.0700
1	18	11.0500	4.7833	4.9750	2.8875
1	19	10.9333	7.7750	5.2875	9.2000

续表

分组	受试对象	细胞培养时间(天)			
		3	7	14	21
1	20	8.9000	11.2000	7.4000	8.9167
2	21	1.3375	0.9750	0.6125	0.0000
2	22	0.2188	0.0857	0.1000	0.0000
2	23	2.3000	2.0833	1.3833	1.7250
2	24	0.9333	0.4667	0.5250	0.4750
2	25	0.6000	0.4750	0.1375	0.0750
2	26	2.0000	0.6000	0.3800	0.3600
2	27	1.4167	0.5625	0.4000	0.3875
2	28	1.0500	0.9375	0.3300	0.2800
2	29	0.1833	0.1500	0.0500	0.0333
2	30	3.9500	1.9125	1.3500	1.0000
2	31	3.7750	3.7667	0.8667	1.5000
2	32	1.3333	2.0500	1.5333	0.4167
2	33	4.3000	3.1375	4.0000	0.3250
2	34	0.4563	0.2643	0.2286	0.1000
2	35	1.9333	3.8000	4.1500	3.2250
2	36	1.8167	0.6333	0.5250	0.5000
2	37	0.8500	1.1500	0.3250	0.5375
2	38	2.8500	3.3000	2.5700	1.1800
2	39	1.2167	0.6625	2.3750	1.3375
2	40	3.8500	3.3250	0.8600	3.3100
2	41	1.1000	1.6200	0.5500	0.2000
2	42	10.0500	4.1750	2.9000	2.9800
2	43	9.7750	7.7000	2.4833	3.1000
2	44	5.2000	3.9333	4.2167	1.7500

三、研究假设

研究问题1:两组眼角膜培养细胞数的总体均数是否相等。
研究问题2:不同时间观察细胞数的总体均数是否相等。
研究问题3:处理和时间有无交互作用。

四、数据录入

1. 变量视图

名称 group　标签 分组
名称 subject　标签 受试对象
名称 time1　标签 3 天
名称 time2　标签 7 天
名称 time3　标签 14 天
名称 time4　标签 21 天

	名称	类型	宽度	小数	标签	值
1	group	数值(N)	8	0	分组	无
2	subject	数值(N)	8	0	受试对象	无
3	time1	数值(N)	8	4	3天	无
4	time2	数值(N)	8	4	7天	无
5	time3	数值(N)	8	4	14天	无
6	time4	数值(N)	8	4	21天	无

2. 数据视图(部分)

	group	subject	time1	time2	time3	time4
1	1	1	.1917	1.6667	1.6500	.7000
2	1	2	2.5938	2.8000	2.4000	3.8000
3	1	3	1.2000	4.3750	4.5250	.2167
4	1	4	2.1792	1.9375	1.6688	.6400
5	1	5	1.0000	1.4500	1.7750	.4250
6	1	6	2.2000	4.7000	3.9667	1.7000
7	1	7	6.6167	5.5667	1.5625	2.5100
8	1	8	6.6000	4.4333	2.4125	1.6875
9	1	9	5.8000	4.6167	2.9875	5.9500
10	1	10	8.0500	7.1000	4.9500	5.0000
11	1	11	.7250	2.1917	2.8000	1.7250
12	1	12	4.4375	6.4563	5.5250	6.9333
13	1	13	3.0625	7.1250	5.1000	.9833

五、操作流程

分析—一般线性模型—重复度量

重复度量定义因子：被试内因子名称：cell

级别数（L）: 4—添加（A）—定义（F）

群体内部变量（W）:3天（time1）对应__?__(1)

7天（time2）对应__?__(2)

14天（time3）对应__?__(3)

21天（time4）对应__?__(4)

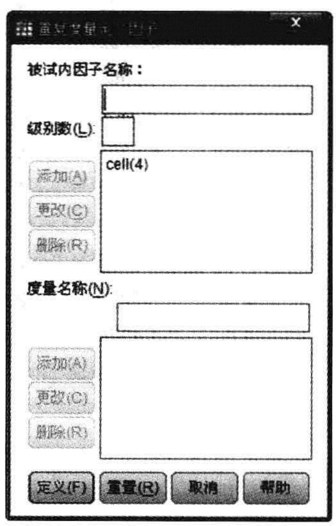
因子列表（B）:分组 group

模型（M）：☑全因子（F）— 继续

选项（O）：☑统计描述（D）— 继续

确定

1. 重复测量需要预先设定组内因素，也称**内因子**，默认的名称为因子1，如下图此处设定为cell，级别数代表重复测量的次数，此处为4，添加之后完成了初步的设定，还需要进一步定义。

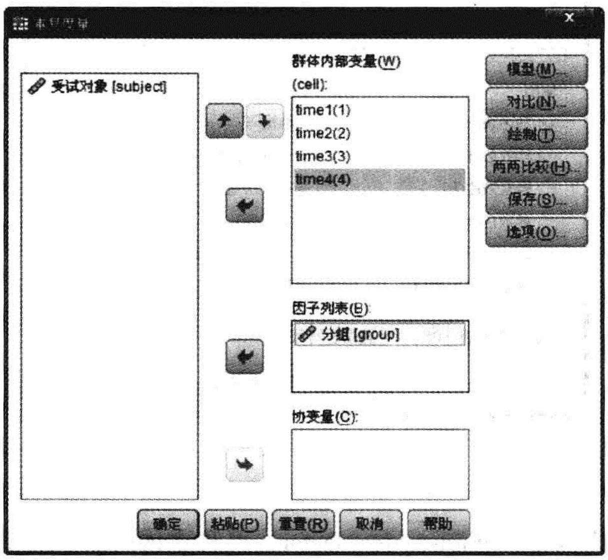


组内因素（within-subject factor）：在重复测量的方差分析模型中，对同一个体相同变量的不同次测量结果被视为一组，称为受试者内因素，如本例当中设定的cell，就包括了time1、time2、time3、time4这样一组变量。可见，组内因素并非在原始数据输入时就产生的，而是在运

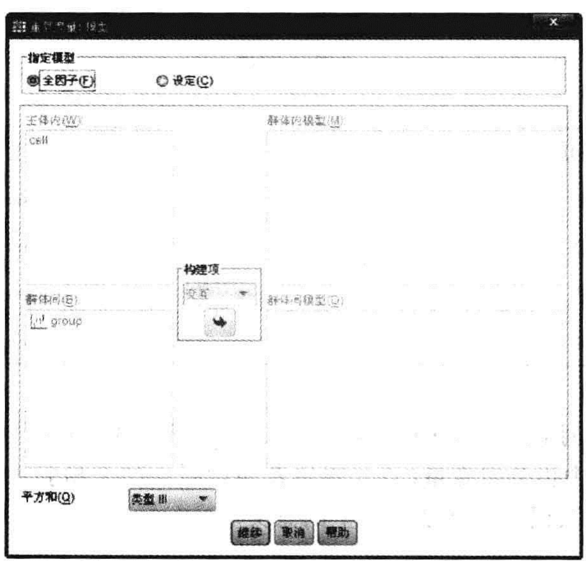
算过程中对一组重复测量的变量进行定义而产生的。

组间因素(between-subject factor):对于受试个体,在重复测量时保持恒定的因素被称为受试者间因素,如本例当中的分组变量 group。

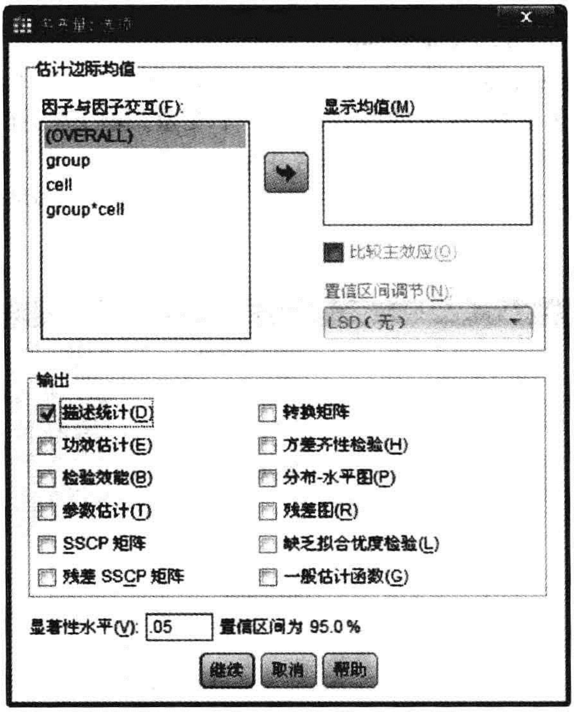
2. 下图为重复测量的方差分析主对话框,我们需要对内因子进行进一步定义,同时可见,由于重复测量的变量为因变量,这里的因变量有 4 个变量,我们需要在群体内部变量(W)中设定,而因子列表(B)相当于自变量,即该研究中的分组变量 group,我们将其选入。可见,重复测量方差分析与其他方差分析的区别在于,由于多次测量而出现多个因变量,因此我们需要对这些因变量进行设定。



3. 我们需要考虑处理因素、时间变量和两者的交互作用,因此为全模型,此处默认即可。



4. 各组的统计描述值是必不可少的,因此选定描述统计(D)。



六、结果解释

1. 下表给出了组内因素(within-subject factor),又称主体内因子的基本情况,通俗来讲,即重复测量的变量,本例指 tim1 ~ time4 的 4 次观察到的平均细胞数。内因子的定义是重复测量分析中的一个关键步骤。

主体内因子

度量:MEASURE_1

cell	因变量
1	time1
2	time2
3	time3
4	time4

2. 下表给出了组间因子(between subject factor),又称主体间因子的基本情况,即试验中感兴趣的因素,本例当中指分组变量(青光眼患者和对照),整个检验中没考虑变量 subject,因为它实际为一个记录 ID,没有任何统计学上的意义。

主体间因子

		N
分组	1	20
	2	24

3. 表中给出了各变量组合下的描述性统计量均数、标准差和例数,大家可按照需要进行挑选。

描述性统计量

	分组	均值	标准 偏差	N
3 天	1	4.775525	3.3609525	20
	2	2.603996	2.6381808	24
	总计	3.591055	3.1480752	44
7 天	1	4.943030	2.7071660	20
	2	1.990242	1.8365822	24
	总计	3.332418	2.6934111	44
14 天	1	3.697400	1.5840600	20
	2	1.368829	1.3556006	24
	总计	2.427270	1.8620658	44
21 天	1	3.416000	2.7909175	20
	2	1.033229	1.1063795	24
	总计	2.116307	2.3530601	44

4. 下表给出了多变量检验结果,相当于将 4 次重复测量视为 4 个因变量而进行多元方差分析,此处意义并不大。

多变量检验[b]

效应		值	F	假设 df	误差 df	Sig.
cell	Pillai 的跟踪	.446	10.745[a]	3.000	40.000	.000
	Wilks 的 Lambda	.554	10.745[a]	3.000	40.000	.000
	Hotelling 的跟踪	.806	10.745[a]	3.000	40.000	.000
	Roy 的最大根	.806	10.745[a]	3.000	40.000	.000
cell * group	Pillai 的跟踪	.082	1.187[a]	3.000	40.000	.327
	Wilks 的 Lambda	.918	1.187[a]	3.000	40.000	.327
	Hotelling 的跟踪	.089	1.187[a]	3.000	40.000	.327
	Roy 的最大根	.089	1.187[a]	3.000	40.000	.327

a. 精确统计量

b. 设计:截距 + group

主体内设计:cell

5. 下表为球形检验(Sphericity Assumption)的结果,以判断各组重复测量的数据之间是否存在相关性。

H_0:重复测量数据之间不存在相关性,数据符合 Huynh-Feldt 条件。

H_1:重复测量数据之间存在相关性,数据不符合 Huynh-Feldt 条件。

a =0.10

所谓 Huynh-Feldt 条件,就是指同一个体的多次重复测量结果间实际不存在相关性,资料的协方差矩阵为 H 型协方差结构。

对于球形检验,通常规定的检验水准 a =0.10,以降低犯Ⅱ类错误的概率。

若 $P>0.10$,则认为数据为球形数据,多次测量结果之间不存在相关性,一般采用单变量方差分析即可;若 $P<0.10$,则认为数据不符合球形数据,多次测量结果之间存在相关性,则采用重复测量的方差分析,校正系数 Epsilon 分别为 Green-house-Geisser、Huynh-Feldt 和下限。本例当中 $P=0.003<0.10$,说明数据不符合球形数据,即不符合 Huynh-Fledt 条件,重复测量数据之间存在相关性,则统计结果需要进行校正。

Mauchly 的球形度检验[b]

度量:MEASURE_1

主体内效应	Mauchly 的 W	近似卡方	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	下限
dimension1 cell	.642	18.075	5	.003	.763	.829	.333

检验零假设,即标准正交转换因变量的误差协方差矩阵与一个单位矩阵成比例

a. 可用于调整显著性平均检验的自由度。在“主体内效应检验”表格中显示修正后的检验

b. 设计:截距 + group

主体内设计: cell

6. 下表为一元方差分析的结果,表中输出的是采用一元方差分析对内因子(本例为 4 个时期)以及内因子与组间因素(group)之间的交互作用有无统计学意义进行检验,并且给出了非校正(采用的球形度)和校正(Greenhouse-Geisser、Huynh-Feldt、下限)的统计结果,三种校正方法推荐采用 Greenhouse-Geisser 的校正结果。本例的球形检验中 $P<0.10$,说明需要进行校正,内因子 cell 校正后的 $F=11.704$,$P<0.01$,说明不同时间观察细胞数的总体均数不全相同;而交互作用 cell * group 经校正后 $F=0.676$,$P>0.05$,说明处理和时间无交互作用。

主体内效应的检验

度量:MEASURE_1

源		Ⅲ型平方和	df	均方	F	Sig.
cell	采用的球形度	65.929	3	21.976	11.704	.000
	Greenhouse-Geisser	65.929	2.290	28.792	11.704	.000
	Huynh-Feldt	65.929	2.487	26.511	11.704	.000
	下限	65.929	1.000	65.929	11.704	.001

续表

源		Ⅲ型平方和	df	均方	F	Sig.
cell * group	采用的球形度	3.810	3	1.270	.676	.568
	Greenhouse-Geisser	3.810	2.290	1.664	.676	.530
	Huynh-Feldt	3.810	2.487	1.532	.676	.542
	下限	3.810	1.000	3.810	.676	.415
误差（cell）	采用的球形度	236.591	126	1.878		
	Greenhouse-Geisser	236.591	96.175	2.460		
	Huynh-Feldt	236.591	104.449	2.265		
	下限	236.591	42.000	5.633		

7. 在重复测量的数据中,除了影响因素外,研究者有时只关心所测量的指标随着时间变化而呈现的变化趋势,而不提供各次测量间两两比较的结果。这里的假设检验为线性、二次或三次曲线拟合成立,故要求检验无统计学意义,以 cell 为例,其线性拟合的 $P<0.01$,二次曲线拟合 $P=0.843$,三次曲线拟合 $P=0.054$,说明二次方曲线拟合最好。

主体内对比的检验

度量:MEASURE_1

源	cell	Ⅲ型平方和	df	均方	F	Sig.
cell	线性	61.959	1	61.959	23.392	.000
	二次	.080	1	.080	.040	.843
	三次	3.891	1	3.891	3.952	.053
cell * group	线性	4.931E-5	1	4.931E-5	.000	.997
	二次	1.442	1	1.442	.721	.401
	三次	2.369	1	2.369	2.406	.128
误差（cell）	线性	111.244	42	2.649		
	二次	83.994	42	2.000		
	三次	41.353	42	.985		

8. 下表给出了组间因素的比较,即青光眼组和对照组的比较,$F=17.843$, $P<0.01$,说明两组眼角膜培养细胞数的总体均数不同。

主体间效应的检验

度量:MEASURE_1

转换的变量:平均值

源	Ⅲ型平方和	df	均方	F	Sig.
截距	1548.506	1	1548.506	104.725	.000
group	263.837	1	263.837	17.843	.000
误差	621.029	42	14.786		

> > 第六章　秩和检验

假设检验的方法可以分为:参数检验(parametric test)和非参数检验(nonparametric test)。

参数检验:若统计推断方法要求样本来自的总体分布已知(如正态分布),在这种假设基础上才能对总体参数(如总体均数)进行估计或检验,因此称为参数检验。但该检验对于分布有严格的要求(正态性、方差齐等),一旦不满足这些条件就非常麻烦,我们常用的 t 检验和方差分析均属于参数检验。

非参数检验:不依赖于特定的总体分布,也不对总体参数进行推断的一类统计方法称为非参数检验,又称任意分布检验(distribution-free test)。

1. 非参数检验适用的资料类型

(1)总体分布非正态或分布形式未知的计量资料(尤其 $n<30$ 时)。

(2)等级资料(见列联表章节)。

(3)个别数据偏大或数据的一端或两端有未确定数值的资料,如 >150 mg。

(4)各组数据离散程度相差悬殊的资料,即方差不齐的资料。

2. 非参数检验的检验效能

非参数检验的适用范围非常广泛,但是由于检验的是分布而不是参数,所以损失了部分资料信息,故通常认为非参数检验方法的检验效能低下。但实际上,根据国外的一项研究,它的检验效能大约为参数检验的 95% 左右,并非低得不能接受,因此当数据可能违背参数检验条件时,最好直接采用相应的非参数检验方法。

3. 非参数检验的方法

非参数检验的方法很多,有符号检验、游程检验、等级相关分析、秩和检验、卡方检验等。本章介绍基于秩次基础上的秩和检验(rank sum test)。

4. 非参数检验分析模块

SPSS 主要在分析下拉菜单项中的非参数检验中实现,其包含的具体统计过程如下:

(1)单样本:单样本秩和检验。

(2)独立样本:两组或多组独立样本秩和检验。

(3)相关样本:配对样本秩和检验。

第一节　单样本秩和检验

一、方法原理

Wilcoxon 符号秩和检验常用于不满足 t 检验条件的单样本连续变量资料的比较,其目的是

推断样本中位数与已知总体中位数(常为标准值或大量观察的稳定值)是否相等。类似于参数检验中的单样本 t 检验,只是因为不是正态,或分布未知,所以不能比较均数,而是比较中位数。

二、分析示例

已知某地正常人尿铅含量的中位数为 2.50 μmol/L。今在该地随机抽取 16 名工人,测定尿铅含量见下表,问该厂工人的尿铅含量是否高于当地正常人?(《卫生统计学》第六版例 10-1,P178)

某厂 16 名工人与当地正常人的尿铅含量(μmol/L)比较

0.62	0.78	2.13	2.48	2.54	2.68	2.73	3.01
3.13	3.27	3.54	4.38	4.38	5.05	6.08	11.27

三、研究问题

H_0:差值的总体中位数等于 0,即该厂工人的尿铅含量与正常人相同。

H_1:差值的总体中位数大于 0,即该厂工人的尿铅含量高于正常人。

单侧 a = 0.05

四、数据录入

1. 变量视图

名称 x 标签 血铅含量

名称	类型	宽度	小数	标签	值
x	数值(N)	8	2	血铅含量	无

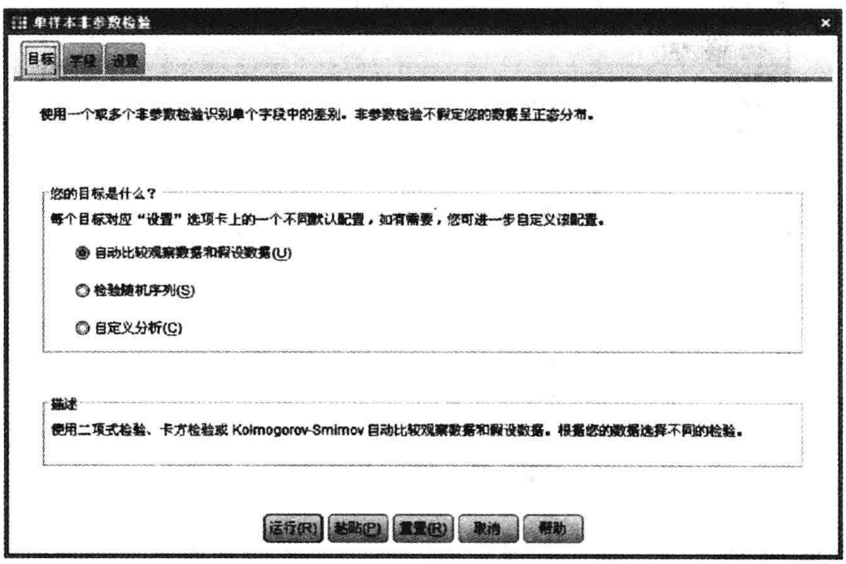
2. 数据视图

	x
1	.62
2	.78
3	2.13
4	2.48
5	2.54
6	2.68
7	2.73
8	3.01
9	3.13
10	3.27
11	3.54
12	4.38
13	4.38
14	5.05
15	6.08
16	11.27

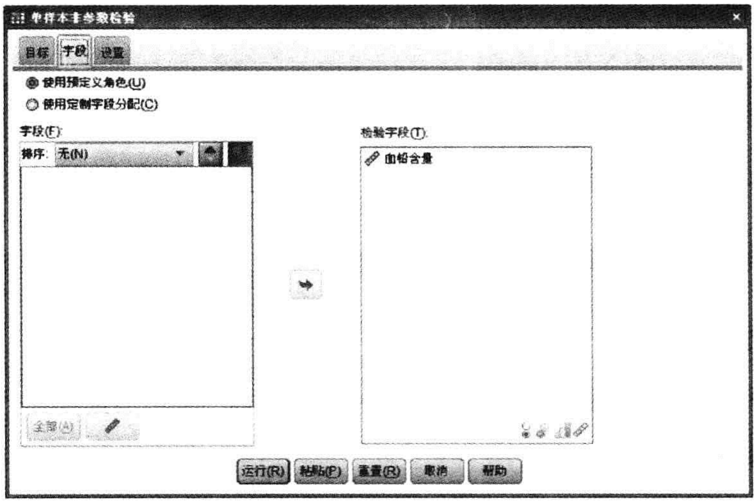
五、操作流程

> 分析—非参数检验—单样本
>
> 目标：自动比较观察数据和假设检验（U）
>
> 字段：检验字段（T）:血铅含量
>
> 设置：自定义检验（T）
>
> 比较中位数和假设中位数（Wilcoxon 符号秩和检验）
>
> 假定中位数：2.5
>
> 运行（R）

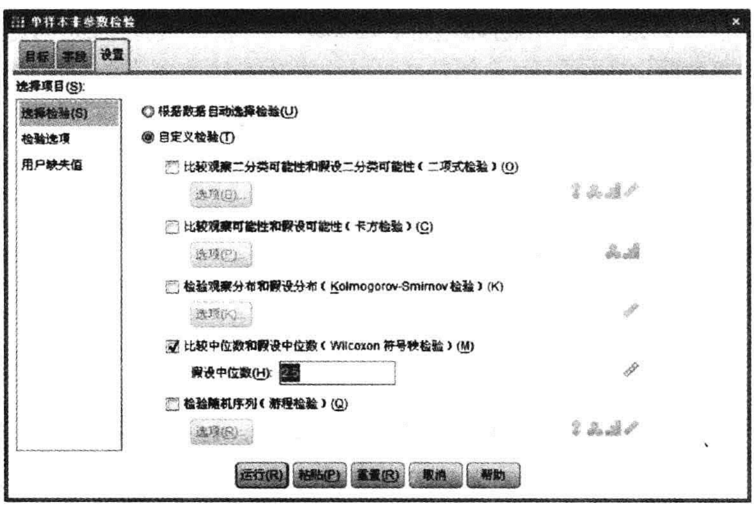
1. 下图指出了非参数检验对于数据的要求：非参数检验不假定您的数据呈正态分布。至于目的，默认为自动比较观察和假设数据，我们先不用管它，因为还需要在其他地方进行设置。



2. 下图为字段选项卡，由于为单变量数据，因此系统自动将分析变量血铅含量 x 选入，不用我们操心。



3. 下图是对检验方法的设定，单样本秩和检验类似于单样本 t 检验，单样本 t 检验是推断样本均数和总体均数是否相等，而单样本秩和检验是推断样本中位数与已知总体中位数是否相等，因此我们选用比较中位数选项，即 Wilcoxon 符号秩和检验，并设定假设中位数 2.5，即比较样本所代表的总体中位数是否与 2.5 相等。



六、结果解释

1. 下表为假设检验摘要，原假设为血铅含量的中位数等于 2.5，检验方法为单样本 Wilcoxon 带符号秩次检验，检验结果为 $P = 0.038 < 0.05$，拒绝原假设，可认为该厂工人的血铅含量高于正常人。

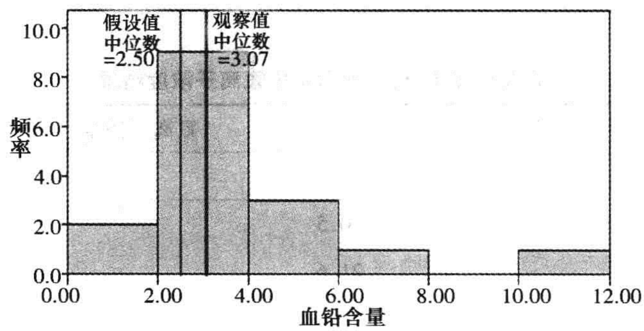
假设检验摘要

	原假设	检验	显著性水平	决策
1	血铅含量的中位数等于2.5	单样本 Wilcoxon 带符号秩次检验	0.038	拒绝原假设

显著渐近显著性。显著性水平为0.05

2. 在结果窗口双击假设检验摘要表格,弹出详细结果窗口,其结果如下:

单样本Wilcoxon 带符号秩次检验



图形横轴显示了血铅含量,纵轴则标出了该含量所对应的频率,同时给出了总体中位数为2.50,而观察值中位数为3.07,至于是否有统计学意义,还需要进一步分析。

3. 下表给出了具体统计量的值为108,也就是正秩和或负秩和的值 T 为 108,$P=0.038<0.05$,注意统计软件标注的为双侧检验,此处可以类推为单侧检验结果。

总计 N	16
检验统计	108.000
标准误	19.326
标准化检验统计	2.070
渐进显著性水平(双侧检验)	0.038

注意:检验水准有单侧和双侧之分,例如,对立假设为两个总体均数不相等,因为 $\mu_1 \neq \mu_2$ 包含 $\mu_1>\mu_2$ 和 $\mu_1<\mu_2$ 两种情形,故称为**双侧检验**;如果凭借专业知识有充分把握可以排除某一侧,对立假设为 $\mu_1>\mu_2$ 或 $\mu_1<\mu_2$,就称为**单侧检验**。为了稳妥起见,一般情况下均采用双侧检验,但是本例结合专业知识,工厂工人的血铅含量不可能低于正常人,因此为单侧检验。

第二节　配对样本秩和检验

一、方法原理

配对设计资料主要是对差值进行分析,通过检验配对样本的差值是否来自中位数为0的总体来判断两个总体中位数是否相等,即两种处理的效应是否相同。其基本思想为:在配对样

本中，由于随机误差的存在，各对数据差值的产生不可避免，假定两种处理的效应相同，则差值的总体分布为对称分布，并且差值的总体中位数为0。若此假设成立，样本差值的正秩和与负秩和应相差不大，均接近 $n(n+1)/4$；当正负秩和相差悬殊，超出抽样误差可解释的范围时，则有理由怀疑该假设，从而拒绝 H_0。

二、分析示例

对11份工业污水测定氟离子浓度（mg/L），每份水样同时采用电极法及分光光度法测定，就总体而言，这两种方法的测定结果有无差别？（《卫生统计学》第六版例10－2，P179）

两法测定11份工业污水中氟离子浓度结果

样品号	氟离子浓度	
	电极法	分光光度法
1	10.5	8.8
2	21.6	18.8
3	14.9	13.5
4	30.2	27.6
5	8.4	9.1
6	7.7	7.0
7	16.4	14.7
8	19.5	17.2
9	127.0	155.0
10	18.7	16.3
11	9.5	9.5

三、研究假设

H_0：差值的总体中位数等于0。

H_1：差值的总体中位数不等于0。

a＝0.05

四、数据录入

1. 变量视图

名称 x1 标签 电极法

名称 x2 标签 分光光度计

名称	类型	宽度	小数	标签	值
x1	数值(N)	8	1	电极法	无
x2	数值(N)	8	1	分光光度计	无

2. 数据视图

	x1	x2
1	10.5	8.8
2	21.6	18.8
3	14.9	13.5
4	30.2	27.6
5	8.4	9.1
6	7.7	7.0
7	16.4	14.7
8	19.5	17.2
9	127.0	155.0
10	18.7	16.3
11	9.5	9.5

五、操作流程

分析—非参数检验—相关样本

目标：自动比较观察数据和假设检验（U）

字段：检验字段（T）:电极法，分光光度计

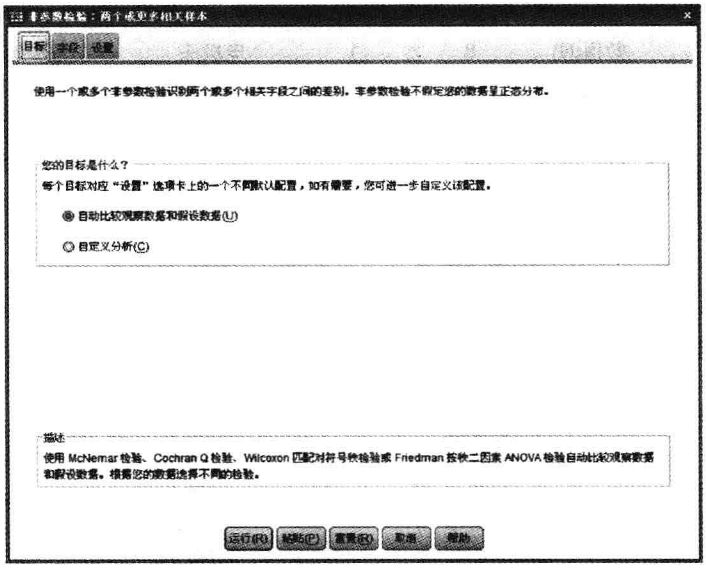
设置：自定义检验（C）

比较中位数和假设中位数差

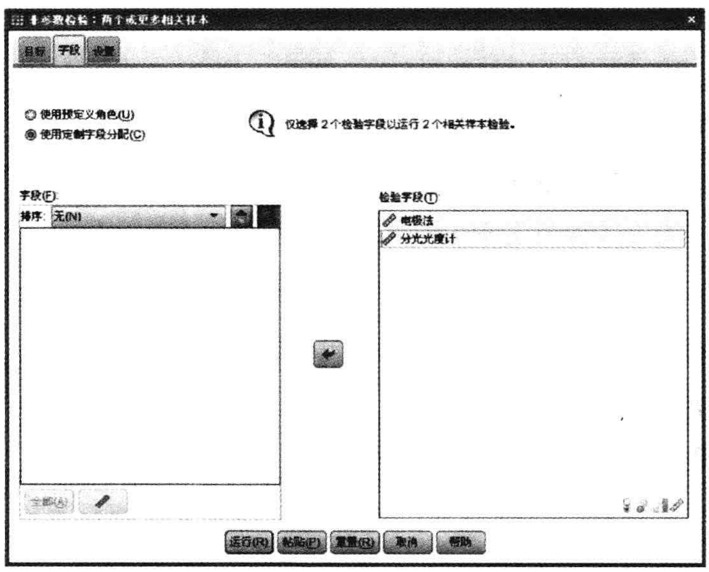
Wilcoxon 匹配对等符号秩（二样本）（W）

运行（R）

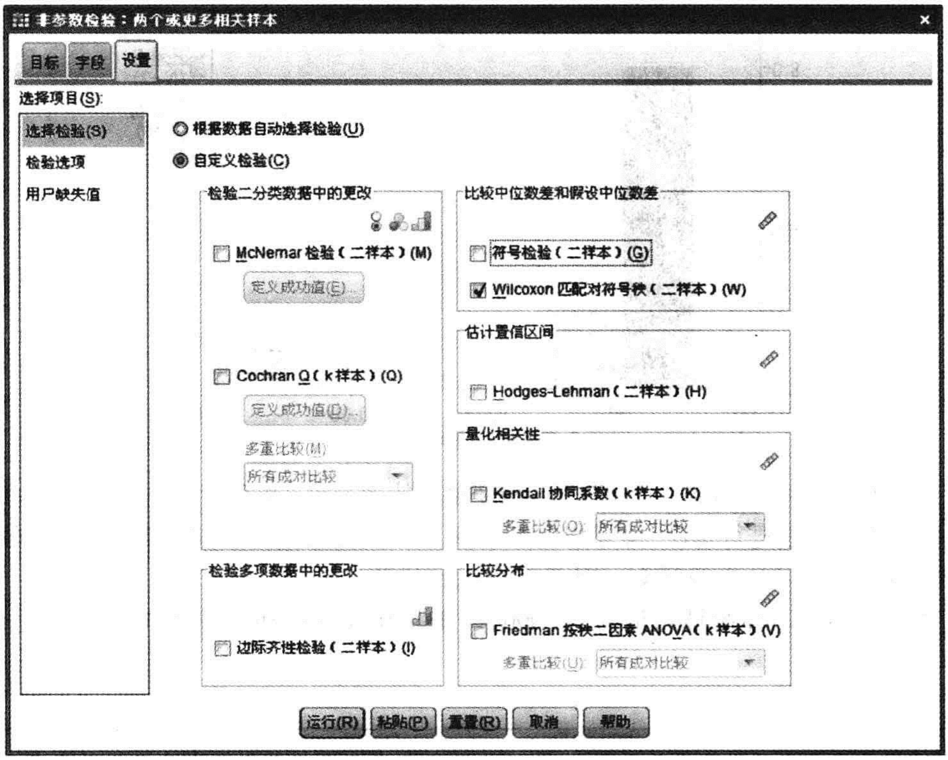
1. 下图指出了非参数检验对于数据的要求:非参数检验不假定数据呈正态分布。至于目的,默认为自动比较观察和假设数据,我们先不用管它,因为还需要在其他地方进行设置。



2. 下图为字段选项卡，配对变量必定选入两个变量，即将变量电极法 x1 和分光光度计 x2 均选入。



3. 下图是对检验方法的设定，研究目的在于比较两个样本差值所代表的总体中位数是否为 0，所对应的配对样本 t 检验则是两样本差值所代表的总体均数是否为 0。而配对样本秩和检验又包括符号检验和 Wilcoxon 配对符号秩，符号检验只利用了正负号，其检验效率低下，而 Wilcoxon 配对符号秩为常用的配对设计差值的秩和检验，它利用了资料的秩序大小，因此采用 Wilcoxon 配对符号秩检验。



六、结果解释

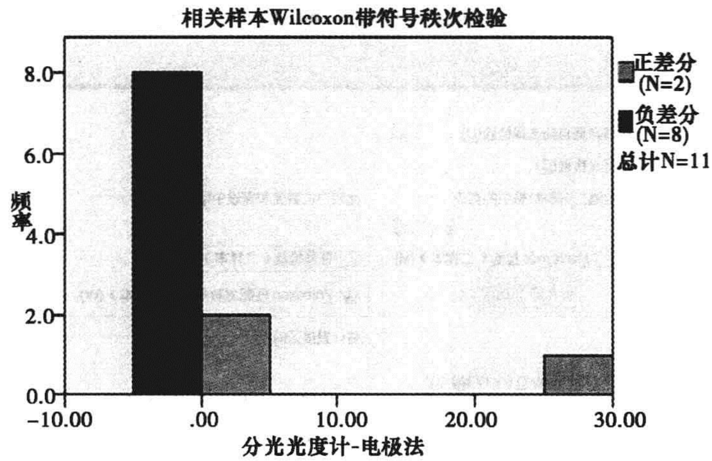
1. 下表对检验结果作了简单介绍，原假设为电极法和分光光度计差值所代表的总体中位数为0，检验方法为相关样本 Wilcoxon 带符号的秩次检验，即配对秩和检验，$P=0.092>0.05$，因此可认为电极法和分光光度计两种方法的测定结果无差别。

假设检验摘要

	原假设	检验	显著性水平	决策
1	电极法和分光光度计间差值的中位数等于0	相关样本 Wilcoxon 带符号秩次检验	0.092	保留原假设

显示渐进显著性。显著性水平为0.05

2. 在结果窗口双击假设检验摘要表格，弹出详细结果窗口，其结果如下：



分光光度计-电极法得到一个新变量,该变量可能为正值,也可能为负值,我们将所有值按照绝对值的大小分别进行排序,得到相应的频率图。从图上可以看出,得到的正值较少($n=2$),而负值较多($n=8$),并且还有一对差值为0,这样共计11对。

3. 下表给出的具体统计量值为11.000,即正秩和或负秩和 T 值为11.000,$P=0.092>0.05$

总计N	11
检验统计	11.000
标准误	9.804
标准化检验统计	-1.683
渐进显著性水平(双侧检验)	0.092

第三节　两组独立样本秩和检验

一、方法原理

Wilcoxon 秩和检验(Wilcoxon rank sum test)目的是推断连续变量资料或有序变量资料的两个独立样本所代表的两个总体分布是否有差别。理论上零假设 H_0为两总体分布相同,即两样本来自同一总体;其对立假设 H_1为两总体分布不同。由于秩和检验对两总体分布形状的差别不敏感,对位置相同、形状不同但类似的两总体分布,推断不出两总体分布形状有差别,故对立的备择假设 H_1不能为两总体分布不同,而只能为两总体分布位置不同。故在实际应用中检验假设 H_0可写作两总体分布位置相同,也可简化为两总体中位数相等。

二、分析示例

用两种药物杀灭钉螺,每批用200~300只活钉螺,用药后清点钉螺的死亡数,并计算死亡率

(%),结果见表,问两种药物杀死钉螺的效果有无差别?(《卫生统计学》第六版例 10-3,P181)

两种药物杀灭钉螺死亡率(%)的比较

甲药死亡率(%)	乙药死亡率(%)
32.5	16.0
35.5	22.5
40.5	26.0
40.5	28.5
49.0	32.5
49.5	38.0
51.5	40.5

三、研究假设

H_0:两种药物杀灭钉螺死亡率的总体中位数相等。

H_1:两种药物杀灭钉螺死亡率的总体中位数不相等。

a = 0.05

四、数据录入

1. 变量视图

<u>名称</u> group <u>标签</u> 药物 <u>值</u> 1 = 甲 2 = 乙

<u>名称</u> x　　<u>标签</u> 死亡率

	名称	类型	宽度	小数	标签	值
1	group	数值(N)	8	0	药物	{1, 甲}...
2	x	数值(N)	8	1	死亡率	无
3						

2. 数据视图

	group	x
1	1	32.5
2	1	35.5
3	1	40.5
4	1	40.5
5	1	49.0
6	1	49.5
7	1	51.5
8	2	16.0
9	2	22.5
10	2	26.0
11	2	28.5
12	2	32.5
13	2	38.0
14	2	40.5

五、操作流程

分析—非参数检验—独立样本

目标：自动比较不同组间的分布

字段：检验字段（T）:死亡率

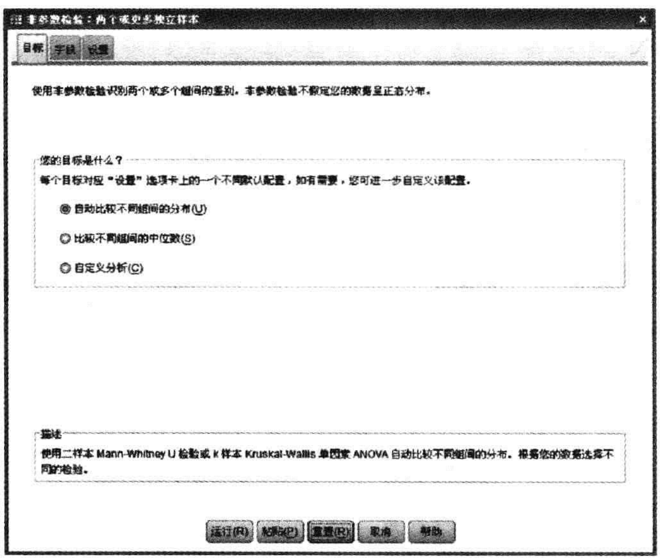
组（G）：药物

设置：自定义检验（C）

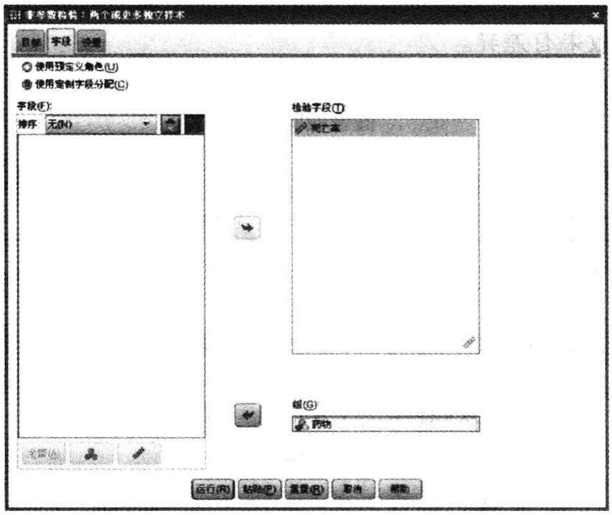
Mann-Whitney U（二样本）（H）

运行（R）

1. 下图指出了非参数检验对于数据的要求:非参数检验不假定您的数据呈正态分布。至于目的,默认为自动比较不同组间的分布,我们先不用管它,因为还需要在其他地方进行设置。



2. 下图为字段选项卡,检验字段(测量指标)框选入死亡率,而组(分组变量)框则选入药物。值得注意的是,有时检验字段的类型必须为数值,组变量的类型必须为字符,变量类型可以在变量视图窗口进行修改。



3. 统计类型采用默认的 Mann-Whitney U,即我们常用的两样本秩和检验,该方法是和成组 t 检验相对应的一种非参数检验方法,在检验时利用了大小次序,即检验 A 样本中的数值是否多数大于 B 样本。此处不应选择中位数检验,该方法检验各个样本是否来自具有相同中位数的总体,但检验效能最低。



六、结果解释

1. 下表对检验结果作了简单介绍,原假设为死亡率的分布在药物的类别间相同,检验方

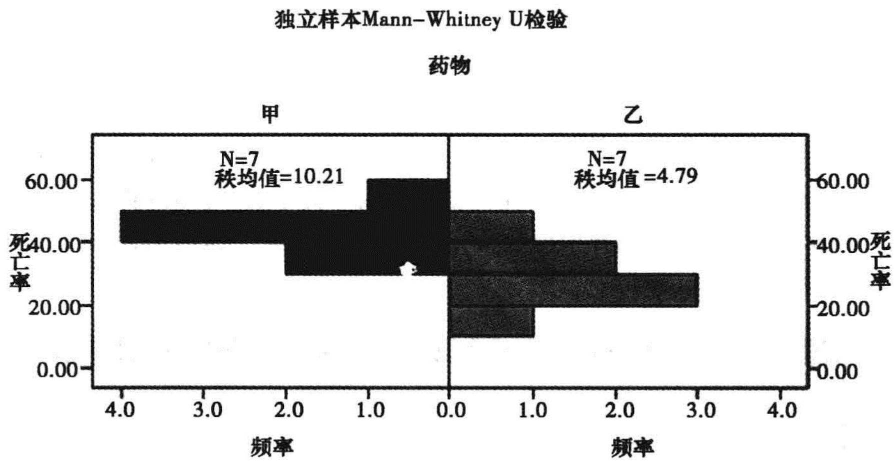
法为 Mann-Whitney，即两组独立样本的秩和检验，其检验结果为 $P=0.015<0.05$，因此可认为两种药物杀死钉螺的效果有差异。

假设检验摘要

	原假设	检验	显著性水平	决策
1	死亡率的分布在药物的类别间相同	独立样本 Mann-Whitney U 检验	0.015	拒绝原假设

显示渐进显著性，显著性水平为 0.05

2. 在结果窗口双击假设检验摘要表格，弹出详细结果窗口，其结果如下：



上图为条图，显示了两组的死亡率的频率分布情况，图上很明显，甲组的死亡率集中于40%，乙组集中于20%，两组差异比较明显，但是具体统计结果如何，还需要进一步的统计分析。

3. 下表为两组秩和检验结果，给出了 Mann-Whitney U 统计量，Wilcoxon W 统计量近似法计算出的 P 值（渐进显著性）和确切概率法计算的 P 值（精确显著性），可见 $P<0.05$，说明两组死亡率的分布差别具有显著性，结合实际数据，可以推断甲药组的死亡率高于乙药组。

总计 N	14
Mann-Whitney U	5.500
Wilcoxon W	33.500
检验统计	5.500
标准误	7.783
标准化检验统计	-2.441
渐进性显著性水平（双侧检验）	0.015
精确显著性水平（双侧检验）	0.011

第四节 多组独立样本秩和检验

一、方法原理

多组独立样本比较的秩和检验是由 Kruskal 和 Wallis 在 Wilcoxon 两样本秩和检验的基础上扩展而来,又称 Kruskal-Wallis H 检验,用于推断连续变量或有序分类变量的多个总体分布有无差别。多组连续变量独立样本的比较,首选单因素方差分析,但是根据经验,本例中生存月数通常不服从正态分布,不满足方差分析的条件,故用 Kruskal-Wallis H 秩和检验。而多组有序分类变量的比较,详见于列联表章节中的单向有序的列联表分析。

二、分析示例

某医院用 3 种不同方法治疗 15 例胰腺癌患者,每种方法各治疗 5 例,治疗后生存月数见下表,问这 3 种方法对胰腺癌患者的疗效有无差别?(《卫生统计学》第六版例 10-5,P183)

3 种方法治疗胰腺癌患者的生存月数比较

甲法	乙法	丙法
3	6	2
4	9	3
7	10	5
8	12	7
8	13	8

三、研究假设

H_0:3 种方法治疗后患者生存月数的总体中位数相等。

H_1:3 种方法治疗后患者生存月数的总体中位数不全相等。

a = 0.05

四、数据录入

1. 变量视图

名称 treat　　标签 治疗方法　值 1 = 甲法 2 = 乙法 3 = 丙法

名称 month　　标签 生存月数

	名称	类型	宽度	小数	标签	值
1	treat	数值(N)	8	0	治疗方法	{1, 甲法}...
2	month	数值(N)	8	0	生存月数	无
3						

2. 数据视图

	treat	month
1	1	3
2	1	4
3	1	7
4	1	8
5	1	8
6	2	6
7	2	9
8	2	10
9	2	12
10	2	13
11	3	2
12	3	3
13	3	5
14	3	7
15	3	8

五、操作流程

分析—非参数检验—独立样本

目标：自动比较不同组间的分布（U）

字段：检验字段（T）:生存月数

组（G）：治疗方法

设置：自定义检验（T）

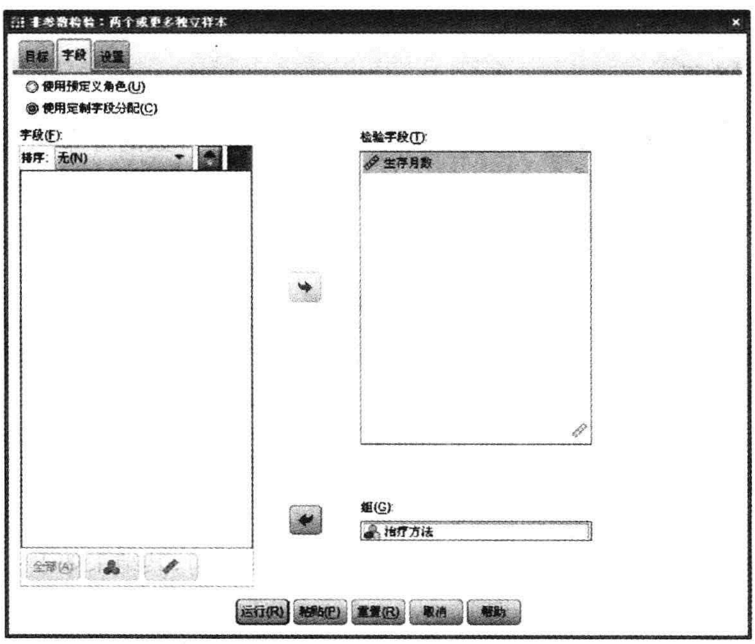
Kruskal-Wallis 单因素 ANOVA（K 样本）（W）

运行（R）

1. 下图指出了非参数检验对于数据的要求：非参数检验不假定您的数据呈正态分布。至于目的，默认为自动比较不同组间的分布，我们先不用管它，因为还需要在其他地方进行设置。



2. 下图为字段选项卡,检验字段(测量指标)框选入生存月数,而组(分组变量)框则选入治疗方法。值得注意的是,有时检验字段的类型必须为数值,组变量的类型必须为字符,变量类型可以在变量视图窗口进行修改。



3. 在选择检验类型中,我们选用 Kruskal-Wallis 单因素 ANOVA(见下图),也就是常用的多样本比较的秩和检验,此处出现了 ANOVA,即方差分析,提示该检验与方差分析存在着某种联系,其实 Kruskal-Wallis 单因素 ANOVA 就是首先将各组数据一起进行排序,将排序的结果

形成新的秩变量,然后对新生成的秩变量进行常规的单因素 ANOVA。我们不应选择中位数检验,它虽然直接检验多个样本所代表的各总体的中位数是否相同,但是检验效能很低,很难发现组间的差别。



六、结果解释

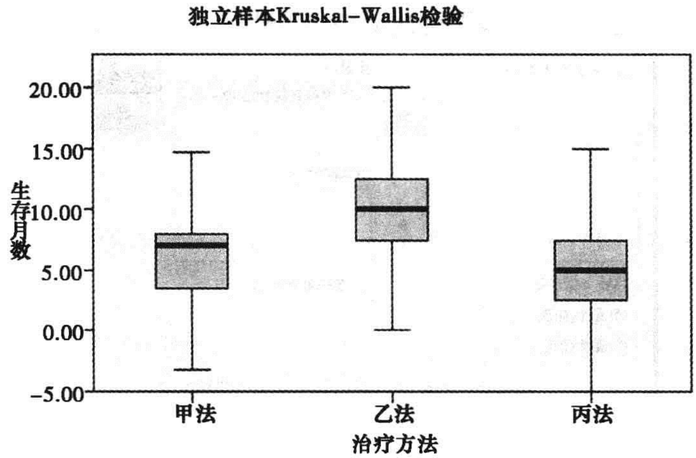
1. 下表对检验结果作了简单介绍,原假设为各组的生存月数的分布相同,检验方法为独立样本 Kruskal-Wallis 检验,即多组独立样本的秩和检验,检验结果为 $P = 0.041 < 0.05$,因此可以认为 3 种治疗的生存月数有差异。

假设检验摘要

	原假设	检验	显著性水平	决策
1	生存月数的分布在治疗方法的类别间相同	独立样本 Kruskal-Wallis 检验	0.041	拒绝原假设

显示渐进显著性。显著性水平为 0.05

2. 在结果窗口双击假设检验摘要表格,弹出详细结果窗口,其结果如下:



这是常见的箱式图（box plot），常用于描述不符合正态分布的连续变量的分布特征，它表现了连续型变量的5个特征值，即最小值、下四分位数（Q_1，即25%百分位数）、中位数（M，Median）、上四分位数（Q_3，即75%百分位数）和最大值。由 Q_1 和 Q_3 构成箱式图的“箱体”部分，箱体的距离为四分位数间距（interquartile range，IQR）。Q_1 和最小值之间、Q_3 和最大值之间分别构成“箱子”的上下两条“触须”，箱体中间的横线为该组的中位数。但该图大家需要特别注意，纵轴所代表的生存月数并非原始数据，而是排序后形成的生存月数秩次，图形能帮助大家直观地感受各组之间的数据分布。

3. 下表给出了具体的统计值，检验统计量校正 H 为6.388，$P=0.041<0.05$。

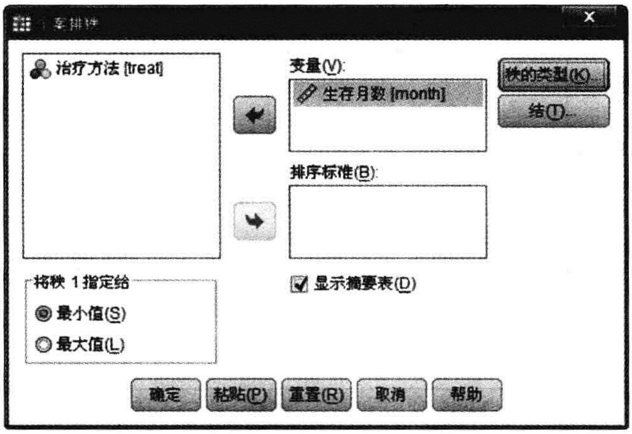
总计 N	15
检验统计	6.388
自由度	2
渐进显著性水平（双侧检验）	0.041

检验统计量已针对结进行调整

七、注意事项

1. 从操作界面可以看出，多组秩和检验只是参数检验中方差分析对非参数分析的一个简单延伸，但是与参数检验比较而言，秩和检验的统计方法较少，如当例题中各组总体出现差异时（$P<0.05$），无法进行两两比较来发现哪些组之间存在差异，因此我们可以对数据进行排秩，然后以秩次作为测量变量，应用参数统计方法进行统计分析更为实用。操作如下：

> 转换（T）—个案排序（K）
>
> 变量（V）：生存月数（month）—确定



2. 操作完成后，结果窗口显示已创建了新变量 Rmonth，因此我们只需对 Rmonth 采用普通的方差分析(ANOVA)即可。

已创建的变量[b]

源变量	函数	新变量	标签
month[a]	秩	Rmonth	Rank of month

a. 秩按升序排列

b. 相同的值的平均秩用于结

>> 第七章　相关分析

相关分析用于描述两个变量间联系的密切程度,它反映的是当控制了其中一个变量的取值后,另一个变量还有多大的变异程度。相关分析有一个显著的特点是变量不分主次,被置于同等地位。线性相关仅适用于二元正态分布,秩相关则适用于不服从二元正态分布、总体分布未知的资料或等级资料。

一、相关分析与回归分析的关系

1. *相关分析*　当两个变量均为结局变量,两变量共同变化,即互相依赖关系,如一个人的身高和体重,用相关分析。

2. *回归分析*　当一个变量依赖于另一个变量,如儿子与父亲的身高有某种依赖关系,用回归分析。

二、SPSS 相关分析模块

SPSS 主要在分析下拉菜单项中的相关中实现,其包含的具体统计过程如下:

1. *双变量(bivariate)*　用于进行两个/多个变量间的参数/非参数相关分析,如果是多变量,则给出两两相关的分析结果。这是最常用的相关分析过程,本章中的线性相关和秩相关均在此过程中进行。

2. *偏相关(partial)*　如果需要进行相关分析的两个变量其取值均受到其他变量的影响,就可以利用偏相关分析对其他变量进行控制,输出控制其他变量影响后的相关系数。

3. *距离(distance)*　此过程可对同一变量内部各观察单位间的数值或各个不同变量间进行相似性或不相似性(距离)分析,前者可用于检测观察值的接近程度,后者则常用于考察各变量的内在联系和结构。该过程一般不单独使用,而是作为因子分析、聚类分析和多维尺度分析的预分析过程,以帮助了解复杂数据集的内在结构,为进一步分析作准备。

第一节　线性相关

一、方法原理

两个随机变量 X,Y 之间呈线性趋势的关系称为线性相关(linear correlation),又称简单相关(simple correlation),其统计学指标为 Pearson 积矩相关系数(Pearson product moment coefficient)。

二、分析示例

某地一项膳食调查中，随机抽取了14名40～60岁的健康妇女，测得每人的基础代谢(kJ/d)与体重(kg)数据，试分析这两项指标间有无关联？(《卫生统计学》第六版 例11-1，P195)

14名中年健康妇女的基础代谢与体重的测量值

编号	基础代谢(kJ/d)	体重(kg)	编号	基础代谢(kJ/d)	体重(kg)
1	4175.6	50.7	8	3970.6	48.6
2	4435.0	53.7	9	3983.2	44.6
3	3460.2	37.1	10	5050.1	58.6
4	4020.8	51.7	11	5355.5	71.0
5	3987.4	47.8	12	4560.6	59.7
6	4970.6	62.8	13	4874.4	62.1
7	5359.7	67.3	14	5029.2	61.5

三、研究假设

研究问题：基础代谢与体重的直线相关系数。

四、数据录入

1. 变量视图

名称 x1 标签 基础代谢

名称 x2 标签 体重

	名称	类型	宽度	小数	标签	值
1	x1	数值(N)	8	1	基础代谢	无
2	x2	数值(N)	8	1	体重	无

2. 数据视图

	x1	x2
1	4175.6	50.7
2	4435.0	53.7
3	3460.2	37.1
4	4020.8	51.7
5	3987.4	47.8
6	4970.6	62.8
7	5359.7	67.3
8	3970.6	48.6
9	3983.2	44.6
10	5050.1	58.6
11	5355.5	71.0
12	4560.6	59.7
13	4874.4	62.1
14	5029.2	61.5

五、操作流程

(一)操作流程(预分析)

图形 — 图形画板模板选择程序

基本：按住 CTRL 键选择 基础代谢,体重 两个变量 — 用鼠标选择右侧 散点图

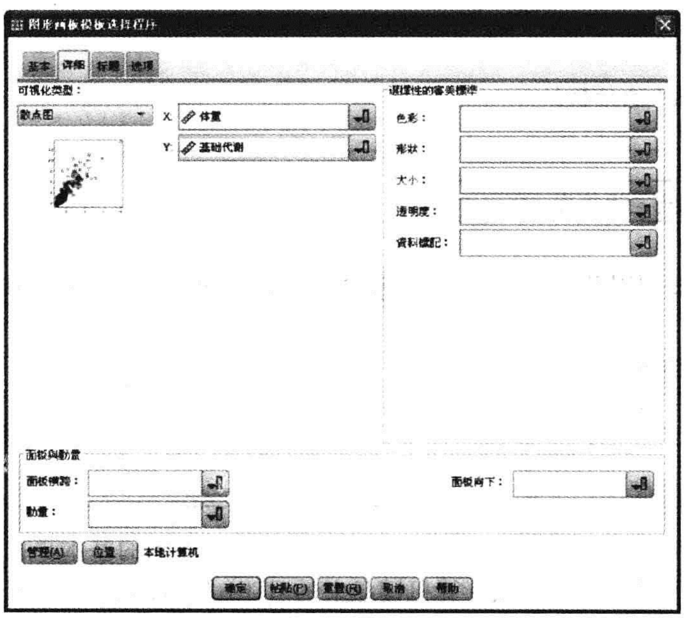
详细：X: 体重 Y: 基础代谢

确定

1. 图形画板模板选择程序是一个傻瓜式的操作界面,下图的基本界面是用来设定入选变量和图形,多个变量的选择需要按计算机键盘上的 CTRL 键,此处同时选择基础代谢和体重两个变量,右侧则选择需要的散点图。



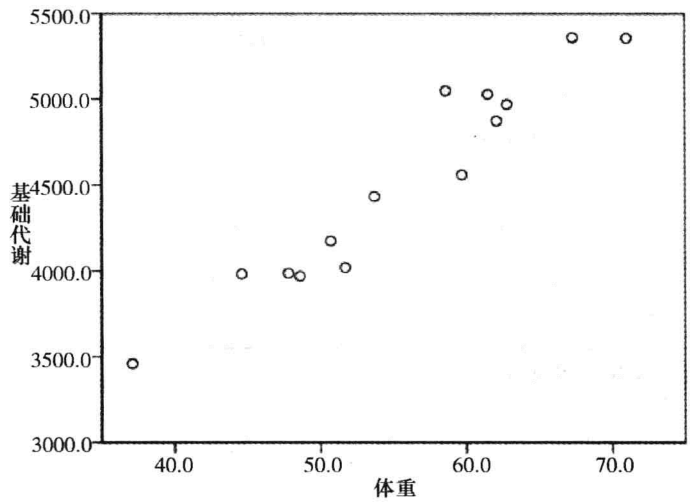
2. 详细对话框需要设定 X 轴和 Y 轴，如果按照默认，图形 X 轴为基础代谢，Y 轴为体重，因此我们需要重新设置，将 X 轴设为体重，Y 轴设为基础代谢。



3. 结果解释

在进行相关分析时，散点图是非常重要的工具，分析前必须做散点图，以初步判断两变量间是否存在相关趋势，该趋势是否为直线趋势，以及数据中是否存在异常点。忽视散点图而直接进行分析可能会得出错误的结论。从下图可以看出，同一基础代谢的人，体重有的要重一

些,而有些又要轻一些,但是从总的趋势来看体重越重,基础代谢越高,反之体重越轻,基础代谢越低,说明基础代谢与体重之间可能存在呈线性趋势的联系,且变化方向相同。



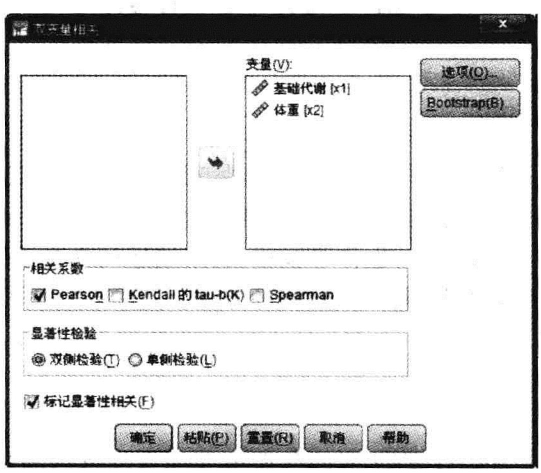
(二)操作流程(正式分析)

分析——相关——双变量 变量(V): 基础代谢 x1 体重 x2 相关系数:☑ Pearson 确定

1. 下图为对话框中相关系数的三个选项,主要体现在对两个变量的要求不同:

(1)Pearson:又称线性相关系数(linear correlation coefficient),是定量表述两个连续变量间线性关系密切程度和相关方向的统计指标。

(2)Kendall 的 tau-b(K)和 Spearman:两者均为等级相关。



2. 结果解释

相关性

		基础代谢	体重
基础代谢	Pearson 相关性	1	.964**
	显著性(双侧)		.000
	N	14	14
体重	Pearson 相关性	.964**	1
	显著性(双侧)	.000	
	N	14	14

注：**. 在0.01 水平(双侧)上显著相关

在上面的结果中，可见 Pearson 相关系数大小为0.964，该数值给出了两个信息，相关性的大小为0.964，相关系数绝对值为0到1之间，值越大，相关性越强；并且给出了相关的方向，由于0.964>0，说明两变量为正相关，即体重越重，则基础代谢越高。对相关系数的双侧检验的 $P<0.01$，说明该相关系数具有统计学意义。

Pearson 相关系数用于描述线性相关，其数值介于 -1 到 1 之间，当两变量相关性达到最大，散点呈一条直线时取值为 -1 或 1，正负号表明了相关的方向；如两变量完全无关，则取值为0。

六、注意事项

1. 进行直线相关分析前，必须先作散点图，以初步判断两变量之间是否存在相关趋势，该趋势是否为线性趋势，以及数据中是否存在异常点。

2. 相关分析不一定是因果关系，如某对夫妇生儿种树，儿长树高，相关关系有统计学意义，可两者并非因果关系，是由于时间变量与两者的潜在联系，造成了身高与树高相关的假象。两变量之间相关系数有统计学意义，欲下因果关系的结论，还需从专业角度作进一步研究。

第二节　秩相关

一、方法原理

Pearson 相关分析仅适用于二元正态分布资料，对那些不服从双变量正态分布的资料，还有总体分布未知的资料和原始数据用等级表示的资料，均不宜用 Pearson 相关系数来描述相关性，可采用秩相关(rank correlation)，也称等级相关，来描述两个变量间相关的程度和方向。这类分析对原变量的分布不作要求，属非参数统计方法。其中最常用的统计量为 Spearman 秩相关系数，又称等级相关系数。

二、分析示例

某研究者观测了10例6个月到7岁的贫血患儿的血红蛋白与贫血体征，研究其相关性，

结果见表,试作秩相关分析。(《卫生统计学》第六版 例 11-4,P201)

贫血患儿的血红蛋白含量(g/dl)和贫血体征

病人编号	1	2	3	4	5	6	7	8	9	10
血红蛋白含量	5.0	5.8	6.1	7.3	8.8	9.1	11.1	12.3	13.5	13.8
贫血体征	+ + +	+ +	+	-	+ +	+ +	-	-	-	-

三、研究假设

研究问题:血红蛋白和贫血体征之间的相关性。

四、数据录入

1. 变量视图

名称 x1 标签 血红蛋白含量

名称 x2 标签 贫血体征

	名称	类型	宽度	小数	标签	值
1	x1	数值(N)	8	1	血红蛋白含量	无
2	x2	数值(N)	8	1	贫血体征	无

2. 数据视图

贫血体征为等级变量,我们需要用数值来表示,0 代表 -,1 代表 +,2 代表 ++,3 代表 +++。

	x1	x2
1	5.0	3.0
2	5.8	2.0
3	6.1	1.0
4	7.3	.0
5	8.8	2.0
6	9.1	2.0
7	11.1	.0
8	12.3	.0
9	13.5	.0
10	13.8	.0

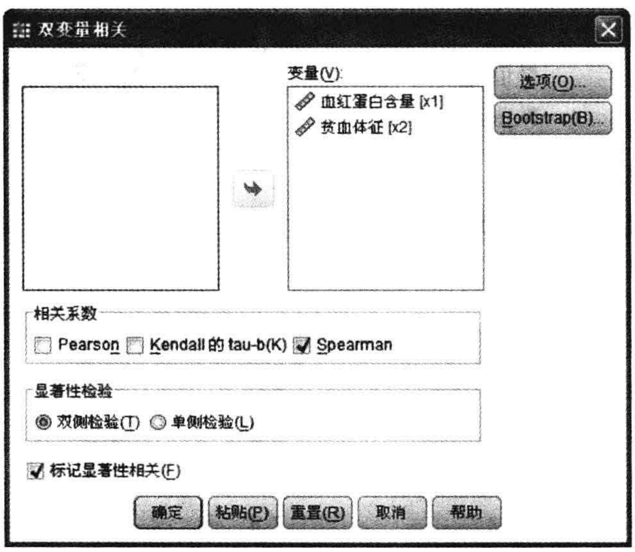
五、操作流程

分析——相关——双变量 变量（V）：血红蛋白含量 x1 贫血体征 x2 相关系数：☑ Spearman 确定

下图对话框中相关系数的三个选项，主要体现在对两个变量的要求不同：

（1）Pearson：又称线性相关系数，是定量描述两个连续变量间线性关系密切程度和相关方向的统计指标。

（2）Kendal 的 tau-b（K）和 Spearman：Spearman 相关系数（秩相关系数）和 Kendall's 等级相关系数均对数据进行排序，但区别在于秩相关是以复合样本整体进行排序，而 Kendall's 等级相关则以每个样本单独进行排序，考察其方向的一致性。Spearman 等级相关系数比较常用。



六、结果解释

在下面的结果中可见，Spearman 相关系数为 -0.741，该数值给出了两个信息，相关性的大小为 0.741，相关系数绝对值大小为 0 到 1 之间，值越大，相关性越强；并且给出了相关的方向，由于 -0.741 <0，说明两变量呈负相关，即血红蛋白含量越高，贫血体征越不明显。对相关系数的双侧检验的 $P=0.014<0.05$，说明该相关系数具有统计学意义。

相关系数

			血红蛋白含量	贫血体征
Spearman 的 rho	血红蛋白含量	相关系数	1.000	-.741*
		Sig.(双侧)	.	.014
		N	10	10
	贫血体征	相关系数	-.741*	1.000
		Sig.(双侧)	.014	.
		N	10	10

注:*. 在置信度(双侧)为0.05 时,相关性是显著的

七、注意事项

1. 对于相同的数据,Spearman 相关系数和 Kendall's 相关系数的绝对值均小于 Pearson 相关系数,显然是由于在秩变换或者数据有序分类处理时损失信息所导致的。

2. Spearman 相关系数的计算公式可以完全套用 Pearson 相关系数的计算公式,但公式中的 X 和 Y 用对应的秩次代替即可。

>>第八章 线性回归分析

一、模型概述

Pearson 积矩相关系数或 Spearman 秩相关系数描述两者之间的关系，重点研究两个变量之间联系的强度和方向，两变量均为结果变量，不分主次。但在实际研究中，我们常需要通过可测或易测的变量对未知或难测的变量进行估计，以达到预测的目的。例如，用患病幼儿月龄预测其体重，从而方便医生制订用药剂量，降低了幼儿因称重带来的不便；用身高或体重这些容易测量的指标来估计体表面积等相对难测的指标等。回归分析(regression analysis)就是研究一个变量如何随另一些变量变化的常用方法。

二、因变量与自变量

被估计或被预测的变量称为因变量(dependent variable)，或称反应变量(response variable)或结果变量，常用 y 表示。y 所依存的变量称为自变量(independent variable)，或称解释变量(explanatory variable)，或称预测因子(predictor)，常用 x 表示。

三、模型解释

因变量只能有一个，自变量可以有一个或多个，若自变量为 p 个，那么 p 个自变量用向量形式表示为($x_1, x_2, \cdots, x_p$)，设有 n 例观察对象，第 i 例($i=1,2,\cdots,n$)的一组观察值为($y_i, x_{i1}, x_{i2}, \cdots, x_{ip}$)。当自变量和因变量之间存在线性关系时，应用线性模型就能很好地刻画它们之间的关系：

$$y_i = \hat{y}_i + e_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip} + e_i$$

可见，实测值 y_i 由两部分组成，第一部分为估计值，用 $\hat{y}$ 表示，读作 y hat，即给定各自变量取值时，因变量 y 的估计值，表示能由自变量决定的部分。第二部分为残差 e_i，是因变量实测值 y 与估计值 $\hat{y}$ 之间的差值，表示不由自变量决定的部分。残差用来判断数据是否符合线性回归分析的条件，判断当前建立的模型是否成立，是否还有别的变量需要引入模型等一系列问题有重要参考价值。

模型如果剔除残差 e_i，就可以得到公式 $\hat{y}_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip}$

b_0为截距(intercept)，在回归方程中又称为常数项(constant)，表示各自变量取值为 0 时 y 的估计值 $b_n x_{in}$ 中 b_n 为偏回归系数(partial regression coefficient)，简称回归系数，表示其他自变量不变，x_{in} 每改变一个单位时，我们所预测的 y 的平均变化量。

第一节　简单线性回归分析

一、方法原理

简单线性回归分析就是寻找因变量数值随自变量的变化而变化的直线趋势，并在散点图上找到一条这样的直线，相应的方程被称为直线回归方程，通过回归方程解释两变量之间的关系更为精确。如果将两事物的取值分别定义为变量 x 和 y，可用回归方程 $\hat{y}=a+bx$ 来描述关系，则有两点需要注意：

1. 变量 x 称为自变量，而 y 为因变量，一般来讲应当有理由认为由于 x 的变化而导致 y 发生变化。

2. $\hat{y}$ 不是一个确定的数值，而是对应于某个确定 x 的群体的 y 值平均值的估计。

二、分析示例

某地一项膳食调查中，随机抽取了 14 名 40～60 岁的健康妇女，测得每人的基础代谢(kJ/d)与体重(kg)数据(见下表)，试拟合直线回归模型？(《卫生统计学》第六版 例 11-1，P195)

14 名中年健康妇女的基础代谢与体重的测量值

编号	基础代谢(kJ/d)	体重(kg)	编号	基础代谢(kJ/d)	体重(kg)
1	4175.6	50.7	8	3970.6	48.6
2	4435.0	53.7	9	3983.2	44.6
3	3460.2	37.1	10	5050.1	58.6
4	4020.8	51.7	11	5355.5	71.0
5	3987.4	47.8	12	4560.6	59.7
6	4970.6	62.8	13	4874.4	62.1
7	5359.7	67.3	14	5029.2	61.5

三、研究假设

研究问题：基础代谢与体重之间的回归分析。

四、数据录入

1. 变量视图

名称 y 标签 基础代谢

名称 x 标签 体重

	名称	类型	宽度	小数	标签	值
1	x1	数值(N)	8	1	基础代谢	无
2	x2	数值(N)	8	1	体重	无

2. 数据视图

	y	x
1	4175.6	50.7
2	4435.0	53.7
3	3460.2	37.1
4	4020.8	51.7
5	3987.4	47.8
6	4970.6	62.8
7	5359.7	67.3
8	3970.6	48.6
9	3983.2	44.6
10	5050.1	58.6
11	5355.5	71.0
12	4560.6	59.7
13	4874.4	62.1
14	5029.2	61.5

五、操作流程

(一)操作流程(预分析)散点图

图形 —图形画板模板选择程序

基本：按住 CTRL 键选择 基础代谢,体重两个变量 — 用鼠标选择右侧 散点图

详细：X: 体重 Y: 基础代谢

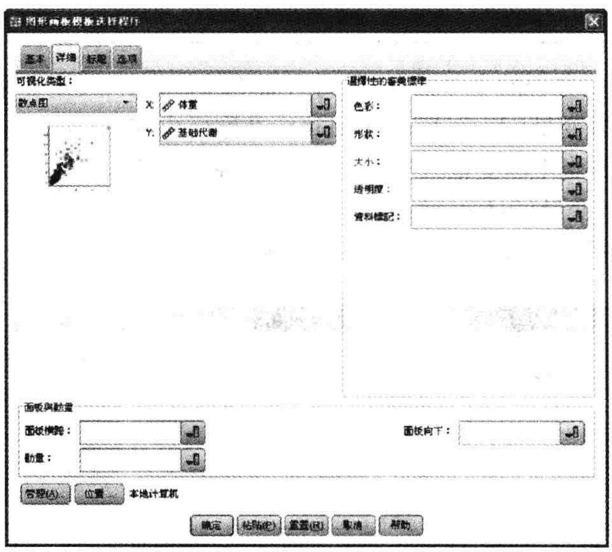
确定

1. 图形画板模板选择程序是一个傻瓜式的操作界面,下图的基本界面是用来设定入选变量和图形,多个变量的选择需要按计算机键盘上的 CTRL 键,同时选择基础代谢和体重两个变

量,右侧则选择需要的图形散点图。

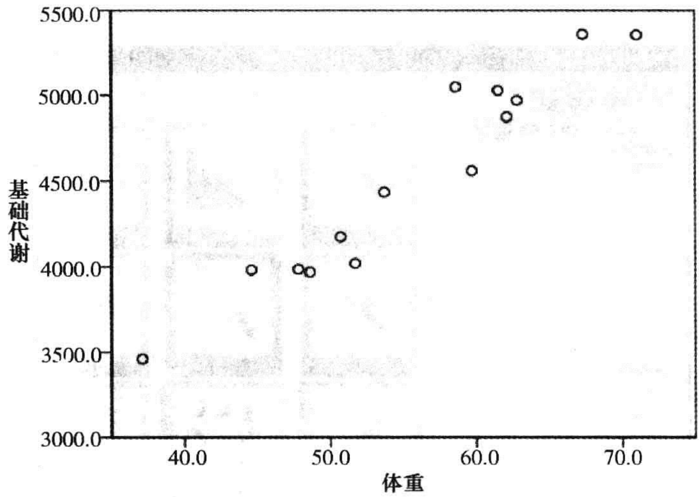


2. “详细”对话框需要设定 X 轴和 Y 轴,如果按照默认,图形默认 X 轴为基础代谢,Y 轴为体重,因此我们需要重新设置,将 X 轴设为体重,Y 轴设为基础代谢。



3. 结果解释

从图中可以看出,因变量 Y 基础代谢和自变量 X 之间呈直线趋势,线性关系成立。



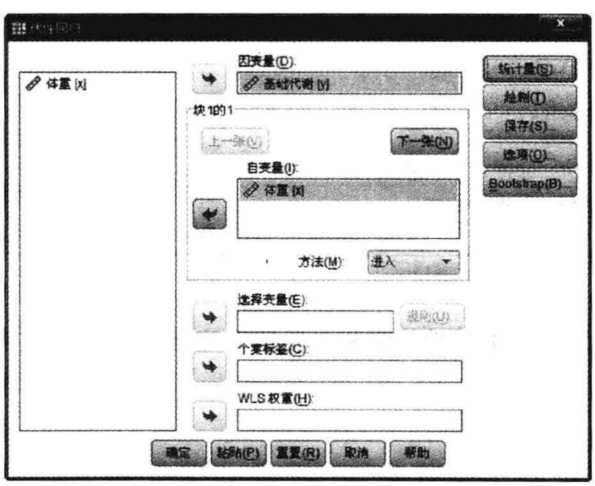
（二）操作流程（线性相关分析）

> 分析——回归——线性
> 因变量(D)框：基础代谢y
> 自变量(I)框：体重x
> 确定

1. 该图为线性回归方程的主对话框，因为为简单回归分析，设定非常简单。

（1）因变量（*D*）：因变量（dependent variable），又称反应变量（response variable），在简单线性回归和多重线性回归当中因变量只有一个，本例指基础代谢。

（2）自变量（*I*）：自变量（independent variable），又称解释变量（explanatory variable）或预测因子（predictor），本例指体重，即体重影响基础代谢。



2. 结果解释

(1)该表格是拟合过程中变量进入/退出模型的情况记录,由于我们只引入了一个自变量,所以只出现了模型1(在多重回归中会依次出现多个回归模型),该模型中体重X为进入变量,没有移出的变量,具体的进入/退出方法为输入(enter),即全部进入模型。如果在多重线性回归中,选择不同的拟合方法则表格内容出现变化,但对于简单线性回归没有多大意义。

输入/移去的变量[b]

模型	输入的变量	移去的变量	方法
1	体重[a]	.	输入

a. 已输入所有请求的变量

b. 因变量:基础代谢

(2)该表为拟合模型的拟合优度情况简报,其重要指标为 R 方(R Square),称为决定系数(coefficient determination),为相关系数的平方。R^2取值在0到1之间,且无单位。它反映了回归贡献的相对程度,即在因变量 Y 的总变异中回归关系所能解释的比例。在实际应用中,通过决定系数反映回归的实际效果。如本例中,$R^2=0.930$,说明40~60岁健康妇女的体重信息大约可以解释自身基础代谢信息量的93%,还有剩余的7%的信息则通过体重以外的其他因素来解释,说明用体重来预测基础代谢量的实际效果较佳。

模型汇总

模型	R	R 方	调整 R 方	标准估计的误差
1	.964[a]	.930	.924	165.1311

a. 预测变量:(常量),体重

(3)该表继续对拟合模型进行检验,其假设检验为总体回归系数 β 是否为0,在本例当中指基础代谢与体重之间是否存在线性关系,其检验方法为方差分析或 t 检验。本例中 $F=158.361$,$P<0.05$,说明 $\beta=0$ 不成立,基础代谢和体重之间存在线性关系。

Anova[b]

模型		平方和	df	均方	F	Sig.
1	回归	4318227.549	1	4318227.549	158.361	.000[a]
	残差	327219.463	12	27268.289		
	总计	4645447.012	13			

a. 预测变量:(常量),体重

b. 因变量:基础代谢

(4)下表为线性回归分析中最重要的一个表格,给出了常数项和变量的系数(包括非标准化系数和标准系数),并对其是否有统计学意义进行检验,一般常数项检验值不用看,t 检验的

结果与前面表格的方差分析结果是一致的。最后得到的回归方程为 $Y = 1106.788 + 61.423X$。

系数[a]

模型		非标准化系数		标准系数	t	Sig.
		β	标准 误差	试用版		
1	(常量)	1106.788	274.534		4.032	.002
	体重	61.423	4.881	.964	12.584	.000

a. 因变量：基础代谢

六、注意事项

1. 进行直线回归分析前，先作散点图，以初步判断两变量之间是否存在回归趋势，该趋势是否为线性趋势，以及数据中是否存在异常点。

2. 相关分析和回归分析具有密切的联系，如果要用统计指标对变量联系的密切程度进行描述，则应当进行相关分析；如果希望反映一个变量变化时对另一个变量的影响大小，则应当使用回归分析。相关系数大小反映了两个变量之间的密切程度，而回归系数反映了 x 和 y 对应的平均数量变化关系，两者的正负号和假设检验是一致的，但两者没有定量的对应关系。

3. 估计值与每个实测值之间的差被称为残差，它刻画了应变量 y 除自身 x 以外的其他所有未进入该模型或未知但可能与 y 有关的随机和非随机因素共同引起的变异，但不能由 x 直接估计的部分。

4. 回归方程中的参数 a 和 b 一般通过最小二乘法原理估计出来，所谓最小二乘法原理就是指使得坐标中每一对 x 变量和 y 变量所对应的点到回归直线纵向距离的平方和，或者说残差平方和最小。

第二节　多重线性回归分析

一、方法原理

多重线性回归(multiple linear regression)是研究一个连续性因变量和多个自变量之间线性关系的统计学分析方法，是简单线性回归的延伸和拓展，其基本原理与简单线性回归一致。

二、分析示例

为了研究有关糖尿病患者体内脂联素水平的影响因素，某医师测定了 30 名患者的体重指数 BMI(kg/m^2)、病程 DY(年)、瘦素 LEP(ng/ml)、空腹血糖 FPG(mmol/L)及脂联素 ADI(ng/ml)水平，数据见下表。(《卫生统计学》第六版 例 13－1，P233)

脂联素水平与相关因素的测量数据

体重指数 (X_1)	病程 (X_2)	瘦素 (X_3)	空腹血糖 (X_4)	脂联素 (Y)	体重指数 (X_1)	病程 (X_2)	瘦素 (X_3)	空腹血糖 (X_4)	脂联素 (Y)
24.22	10.0	5.75	13.6	29.36	24.14	5.0	10.21	7.4	16.01
24.22	3.0	9.32	6.2	14.31	26.45	4.0	19.31	5.1	19.03
19.03	15.0	2.50	11.1	26.08	25.22	2.3	8.65	7.6	17.46
23.39	3.0	5.66	9.7	19.62	27.22	3.0	8.54	8.6	20.36
19.49	4.0	2.83	7.3	42.82	25.93	6.0	7.21	8.9	15.92
24.38	6.0	6.86	7.3	22.76	26.99	12.0	8.75	7.0	15.34
19.03	2.9	3.22	7.7	31.00	25.71	7.0	13.07	13.5	8.05
21.11	9.0	4.90	6.0	17.28	28.41	4.0	8.90	13.5	12.31
23.32	5.0	3.54	6.7	30.25	26.39	4.0	23.26	8.2	5.59
24.34	2.0	4.51	7.2	24.28	28.73	10.0	19.05	6.9	8.59
23.82	8.0	8.47	9.1	18.94	27.46	16.0	19.44	6.5	8.89
22.86	20.0	9.92	8.1	16.08	27.99	10.0	17.33	6.1	14.10
24.49	12.0	6.01	7.0	29.50	28.41	2.0	14.59	6.8	11.74
23.37	6.0	4.31	6.3	25.64	30.69	1.5	22.06	8.1	5.18
20.81	7.0	3.46	7.1	32.26	29.39	3.0	20.56	7.5	6.12

三、研究假设

研究问题：研究脂联素水平 Y 与相关因素（体重指数 X_1、病程 X_2、瘦素 X_3、空腹血糖 X_4）之间的关系。

四、数据录入

1. 变量视图

名称 X1 标签 体重指数
名称 X2 标签 病程
名称 X3 标签 瘦素
名称 X4 标签 空腹血糖
名称 Y 标签 脂联素

	名称	类型	宽度	小数	标签	值
1	X1	数值(N)	8	2	体重指数	无
2	X2	数值(N)	8	2	病程	无
3	X3	数值(N)	8	2	瘦素	无
4	X4	数值(N)	8	2	空腹血糖	无
5	Y	数值(N)	8	2	脂联素	无

2. 数据视图(部分)

	X1	X2	X3	X4	Y
1	24.22	10.00	5.75	13.60	29.36
2	24.22	3.00	9.32	6.20	14.31
3	19.03	15.00	2.50	11.10	26.08
4	23.39	3.00	5.66	9.70	19.62
5	19.49	4.00	2.83	7.30	42.82
6	24.38	6.00	6.86	7.30	22.76
7	19.03	2.90	3.22	7.70	31.00
8	21.11	9.00	4.90	6.00	17.28
9	23.32	5.00	3.54	6.70	30.25
10	24.34	2.00	4.51	7.20	24.28
11	23.82	8.00	8.47	9.10	18.94
12	22.86	20.00	9.92	8.10	16.08

五、操作流程

分析——回归——线性回归

因变量(D):脂联素Y

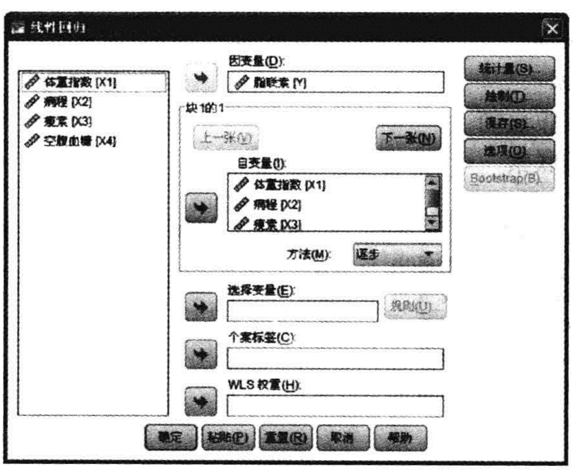
自变量(I):体重指数X1,病程X2,瘦素X3,空腹血糖X4

方法(M):逐步

绘制(T) 直方图(H)正态概率图(R)—继续

确定

1. 该图为线性回归方程的主对话框,设定与简单线性回归类似。



(1)因变量(D):因变量(dependent variable),又称反应变量(response variable),在简单线性回归和多重回归当中因变量只有一个,本例指脂联素 Y。

(2)自变量(I):自变量(independent variable),又称解释变量(explanatory variable)或预测因子(predictor),本例指体重指数 X_1、病程 X_2、瘦素 X_3和空腹血糖 X_4四个变量。

(3)方法(M):多重线性回归有进入、逐步、删除、向前和向后,本例选择逐步法。下面重点介绍两种常用的方法,即进入法和逐步法。

①进入法(enter):即候选自变量全部纳入模型,不作任何筛选,进入法为默认选项。进入法的结果使所有候选变量的 P 值均显示出来,当然也包括无统计学意义的变量,仅适用于自变量个数不太多的情况。

②逐步法(stepwise):又称逐步回归法,对自变量按照一定的纳入和排除标准反复进行引入、剔除过程,直到没有变量被引入,也没有变量被剔除为止。从实际运用上说,逐步法运用最广。

2. 下图为线性回归图形绘制,我们先来谈谈线性回归模型的四个基本条件。

(1)线性趋势:自变量和因变量的关系是线性的,如果不是,则不能采用线性回归来分析,这可以通过散点图来判断。

(2)独立性:可表述为因变量 Y 的取值互相独立,它们之间没有联系。反映到模型中,实际上就要求残差间相互独立,不存在自相关。

(3)正态性:就自变量的任何一个线性组合,因变量 Y 服从正态分布,实际要求残差服从正态分布。

(4)方差齐性:就自变量的任何一个线性组合,因变量 Y 的方差均相同,实质要求残差的方差齐。

故残差分析是回归诊断的一个重要环节,考察残差是否服从正态分布可以通过绘制标准化残差的直方图和正态概率图(PP 图)进行,因变量与自变量间关系非线性、残差方差不齐、观察值间不独立等情况均会导致残差的直方图和正态概率图表现出非正态。



六、结果解释

1. 下面的表格依次列出了模型筛选过程，模型 1 引入了变量瘦素 X_3，而模型 2 引入了变量体重指数 X_1，另外两个变量均未达到进入标准，最终没有纳入。右侧注明的方法为步进法（stepwise），即逐步回归法，其纳入标准为≤0.050，排除标准为≥0.100。

输入/移去的变量[a]

模型	输入的变量	移去的变量	方法
1	瘦素	.	步进（准则：F-to-enter 的概率 ≤.050，F-to-remove 的概率 ≥.100）
2	体重指数	.	步进（准则：F-to-enter 的概率 ≤.050，F-to-remove 的概率 ≥.100）

a. 因变量：脂联素

2. 下表为拟合模型的拟合优度情况简报，其重要指标为 R 方（R Squave），称为决定系数（coefficient determination），为相关系数的平方。R^2 取值在 0 到 1 之间，且无单位。它反映了回归贡献的相对程度，即在因变量 Y 的总变异中回归关系所解释的比例。在实际应用中，通过决定系数反映回归的实际效果。如模型 1 含有变量瘦素，$R^2=0.657$，说明瘦素可以解释脂联素信息的 65.7%，而模型 2 含有变量瘦素和体重指数，说明这两个变量可解释脂联素信息的 70.7%。可见，从上至下随着新变量的引入，模型可解释的变异占总变异的比例越来越大。

模型汇总[c]

模型	R	R 方	调整 R 方	标准估计的误差
1	.811[a]	.657	.645	5.45091
2	.841[b]	.707	.685	5.13104

a. 预测变量：（常量），瘦素

b. 预测变量：（常量），瘦素，体重指数

c. 因变量：脂联素

3. 该表继续对各拟合模型进行检验，即两个模型是否有统计学意义，模型 1 中 $F=53.626$，$P<0.01$，模型 2 中 $F=32.560$，$P<0.01$，可见两个模型均有统计学意义。

Anova[c]

模型		平方和	df	均方	F	Sig.
1	回归	1593.353	1	1593.353	53.626	.000[a]
	残差	831.948	28	29.712		
	总计	2425.301	29			

续表

模型		平方和	df	均方	F	Sig.
2	回归	1714.458	2	857.229	32.560	.000[b]
	残差	710.844	27	26.328		
	总计	2425.301	29			

a. 预测变量：(常量)，瘦素

b. 预测变量：(常量)，瘦素，体重指数

c. 因变量：脂联素

4. 这是线性回归分析中最重要的一个表格，给出了模型 1 和模型 2 的常数项和各变量的系数(包括非标准化系数和标准系数)，并对其是否有统计学意义进行检验。模型 2 为最终的拟合结果，变量瘦素 $t=-3.112$，$P<0.01$；体重指数 $t=-2.145$，$P<0.05$，说明变量各项的偏回归系数均有统计学意义，最后的回归方程为 $Y=53.481-0.753X_3-1.087X_1$

系数[a]

模型		非标准化系数		标准系数	t	Sig.
		β	标准误差	试用版		
1	(常量)	30.528	1.882		16.219	.000
	瘦素	-1.161	.159	-.811	-7.323	.000
2	(常量)	53.481	10.848		4.930	.000
	瘦素	-.753	.242	-.525	-3.112	.004
	体重指数	-1.087	.507	-.362	-2.145	.041

a. 因变量：脂联素

5. 下表反映了没有进入模型的各个变量的检验结果，模型 1 中未引入模型的体重指数 X_1 有统计学意义，说明模型需要继续拟合；模型 2 中未引入模型的各变量均无统计学意义，因此模型不需要继续进行拟合，模型 2 为最终模型。

已排除的变量[c]

模型		Beta In	t	Sig.	偏相关	共线性统计量
						容差
1	体重指数	-.362[a]	-2.145	.041	-.382	.381
	病程	-.027[a]	-.237	.814	-.046	.998
	空腹血糖	-.185[a]	-1.707	.099	-.312	.974
2	病程	-.082[b]	-.756	.457	-.147	.947
	空腹血糖	-.147[b]	-1.385	.178	-.262	.937

a. 模型中的预测变量：(常量)，瘦素

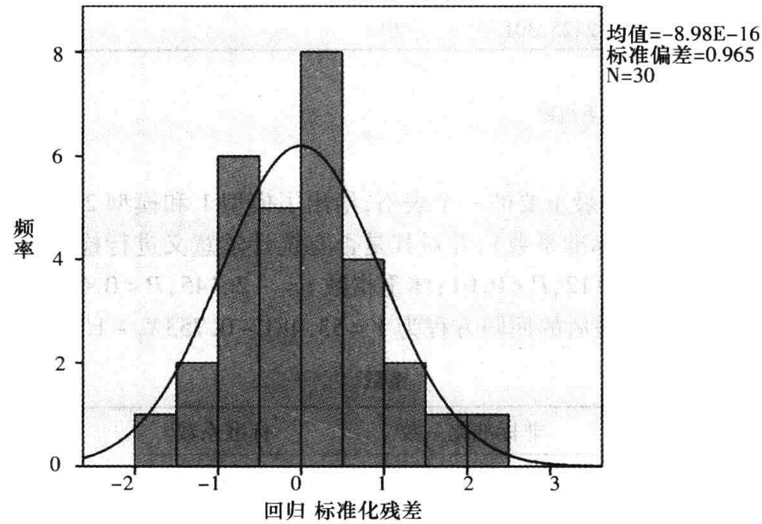
b. 模型中的预测变量：(常量)，瘦素，体重指数

c. 因变量：脂联素

6. 这是残差分析的直方图,图中的曲线为正态参考曲线,可见残差基本呈正态分布,说明该数据比较符合线性回归模型的适用条件,如独立性、正态性和方差齐性。

直方图

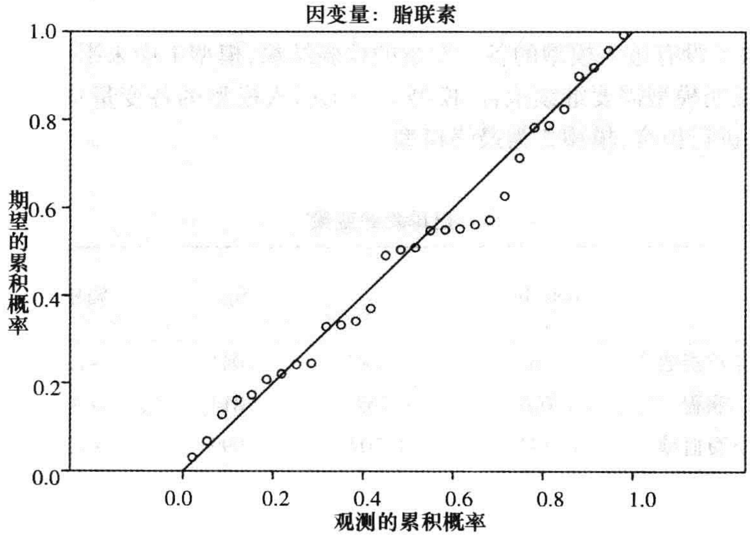
因变量：脂联素



7. 下图为因变量观察累计概率和模型预测值累计概率间的正态 PP 图,用于观察残差分布是否呈正态,可见散点基本围绕参考直线均匀分布,说明该数据比较符合线性回归模型的适用条件,如独立性、正态性和方差齐性。

回归标准化残差的标准 P-P图

因变量：脂联素



>>第九章　四格表卡方检验

一、χ^2检验的基本思想

χ^2检验是判断实际频数和理论已知频数的差别是否由抽样误差引起的，χ^2值的大小反映了实际频数与已知理论频数的吻合程度，在 H_0成立的条件下，二者相差不应该太大，即 χ^2值不大。若算得χ^2值很大，超过了事先设定的检验水准所对应的χ^2界值($\chi^2 \geqslant \chi^2_{0.05}$)，则 $P<0.05$，说明二者的较大差别是由抽样误差引起的可能性很小，则拒绝 H_0，接受 H_1。

二、χ^2检验的用途

1. 频数分布的拟合优度检验。
2. 推断两个变量或特征之间有无关联性。
3. 推断完全随机设计的两组或多组频数分布的概率是否相同。
4. 推断配对设计的两组频数分布的概率是否相同。

三、SPSS 卡方分析模块

1. 拟合优度检验在分析下拉菜单中非参数检验中的单样本非参数检验中实现。
2. 关联性分析和四格表卡方检验均在分析下拉菜单中描述统计中的交叉表中实现。

交叉表(cross table)：是多个频数表的重组，表格中每个格子为列表变量特定值的组合。交叉列表可以检验属于多个变量的观察对象的频数。通过观察频数，我们可以辨别交叉列表中变量间的关系。2×2 表(四格表)是最为简单的交叉列表，只有两个变量，每个变量只有两个特定值。

第一节　一般四格表卡方检验

一、方法原理

一般四格表χ^2检验可以检验两个样本的总体分布是否相同，或者两个样本是否来自同一总体。

二、分析示例

将病情相似的 169 名消化道溃疡患者随机分成两组，分别用洛赛克与雷尼替丁两种药物治疗，4 周后疗效见下表。问两种药物治疗消化道溃疡的愈合率有无差别？(《卫生统计学》第

六版例 9-2，P159）

两种药物治疗消化道溃疡 4 周后疗效

处理	愈合（疗效=1）	未愈合（疗效=2）	合计
洛赛克（药物=1）	64	21	85
雷尼替（药物=2）	51	33	84
合计	115	54	169

三、研究假设

H_0：两药治疗消化道溃疡的愈合率相同，$\pi_1=\pi_2$。

H_1：两药治疗消化道溃疡的愈合率不同，$\pi_1 \neq \pi_2$。

$\alpha=0.05$

四、数据录入

1. 变量视图

名称 treat　标签 处理　值 1 = 洛赛克 2 = 雷尼替丁

名称 result　标签 疗效　值 1 = 愈合 2 = 未愈合

名称 weight　标签 权重

	名称	类型	宽度	小数	标签	值
1	treat	数值(N)	8	0	处理	{1，洛赛克}...
2	result	数值(N)	8	0	疗效	{1，愈合}...
3	weight	数值(N)	8	0	权重	无

2. 数据视图　注意该数据视图和四格表的差别。

	treat	result	weight
1	1	1	64
2	1	2	21
3	2	1	51
4	2	2	33

五、操作流程

数据—加权个案(W)

频率变量(F):权重weight—确定

分析—描述统计—交叉表(C)

行(S):处理treat　　列(C):疗效result

统计量(S)—☑卡方(H)—继续

单元格(E)—计数 ☑观察值

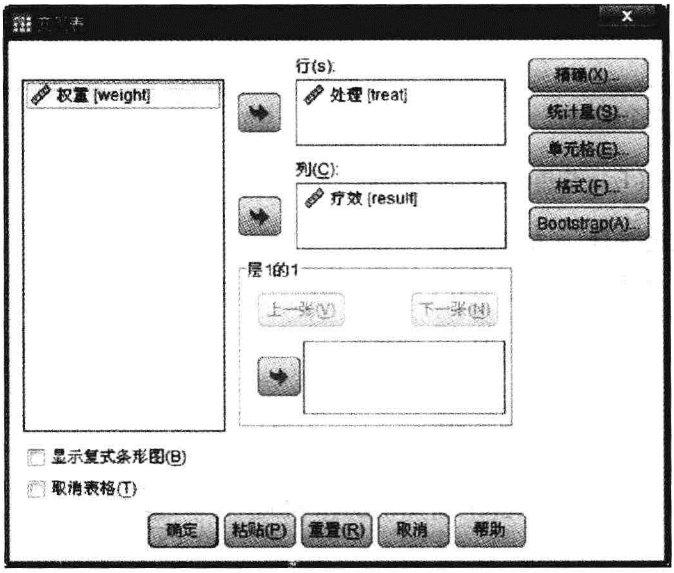
百分比(C) ☑行(R) ☑列(C) ☑总计 —继续

确定

1. 下图是对数据的预定义,在数据(D)下拉菜单框中完成,即设置权重变量,该操作是四格表运算的基本操作。



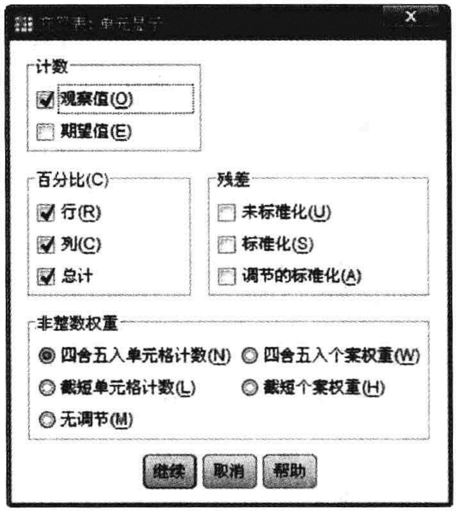
2. 下图是四格表统计的主对话框,主要设置行变量和列变量,该例中行变量指分组变量处理 treat,列变量指结果变量疗效 result,行变量和列变量的设置只对结果的排列产生影响,并不影响其统计结果,因此大家将行变量与列变量调换也可。



3. 下图是对检验方法的选择，选用卡方检验（Chi-square），该检验用于判断行、列变量是否独立，如果数据不满足卡方检验的要求（不能有单元格的期望值 <1，不能有 20% 以上单元格的期望值 <5），则系统会在分析结果的最后给出警告，提示作者采用确切概率法。



4. 下图中各选项是计算四格表中统计描述指标，观察值（observed）指原始数据，期望值（expected）指统计过程中的理论数，一般用处不大。各种百分比，如行百分比（Row）、列百分比（Column）及合计百分比（Total），均为四格表的统计描述内容，比较重要。



六、结果解释

1. 下表是处理记录缺失值情况报告，可见 169 例均为有效值。

案例处理摘要

	案例					
	有效		缺失		合计	
	N	百分比	N	百分比	N	百分比
处理 * 疗效	169	100.0%	0	.0%	169	100.0%

2. 下表标明了各百分比结果，百分数的含义关键是确定分子和分母，分子好确定，也就是各单元格的观察值(计数)，分母则需要确定所对应的 100% 项，如洛赛克组中有三个百分数值，处理中的% 值为 75.3%，对应的 100% 为处理横向合计，即在洛赛克处理组中，治愈率为 75.3%；疗效中% 值为 55.7%，对应的 100% 为纵向疗效值，即在所有愈合的病例中，由洛赛克治疗而愈合的占 55.7%；总计 37.9%，对应的 100% 为右下角的总合计值，即洛赛克治疗愈合的病例占总病例的 37.9%。在本例中，最重要的百分数为处理中%，当然，根据需要，大家可以挑选不同的% 作为统计描述值。

处理 * 疗效 交叉制表

			疗效		合计
			愈合	未愈合	
处理	洛赛克	计数	64	21	85
		处理 中的 %	75.3%	24.7%	100.0%
		疗效 中的 %	55.7%	38.9%	50.3%

续表

			疗效		合计
			愈合	未愈合	
		总数的 %	37.9%	12.4%	50.3%
	雷尼替丁	计数	51	33	84
		处理 中的 %	60.7%	39.3%	100.0%
		疗效 中的 %	44.3%	61.1%	49.7%
		总数的 %	30.2%	19.5%	49.7%
合计		计数	115	54	169
		处理 中的 %	68.0%	32.0%	100.0%
		疗效 中的 %	100.0%	100.0%	100.0%
		总数的 %	68.0%	32.0%	100.0%

3. 下表给出了一堆检验结果,如何选择统计结果,教科书给出了如下判断标准:

(1)当 $n \geqslant 40$ 且所有 $T \geqslant 5$ 时,用 Pearson 卡方,若所得 $P \approx 0.05$ 时,用 Fisher 精确检验。

(2)当 $n \geqslant 40$ 但有 $1 \leqslant T \leqslant 5$ 时,用连续校正卡方。

(3)当 $n < 40$ 或 $T < 1$ 时,用 Fisher 精确检验。

卡方检验

	值	df	渐进 Sig.(双侧)	精确 Sig.(双侧)	精确 Sig.(单侧)
Pearson 卡方	4.130[a]	1	.042		
连续校正[b]	3.487	1	.062		
似然比	4.156	1	.041		
Fisher 的精确检验				.049	.031
线性和线性组合	4.106	1	.043		
有效案例中的 N	169				

a. 0 单元格(.0%)的期望计数<5,最小期望计数为 26.84

b. 仅对 2×2 表计算

本例 $N = 169 \geqslant 40$,且表下 a 注释中说明 0 单元格的期望计数(理论数,T) <5,即 $T \geqslant 5$ 成立,符合条件 1,选用 Pearson 卡方值,$\chi^2 = 4.130$, $P = 0.042 < 0.05$,说明两样本频率的差异具有统计学意义。结合前表中的百分数,可知洛赛克的愈合率为 75.3%,雷尼替丁的愈合率为 60.7%,可以认为洛赛克治疗的愈合率比雷尼替丁的愈合率高。

七、注意事项

1. 连续校正卡方只在四格表中才计算。

2. 研究表明，在 Pearson 卡方、似然比卡方和 Fisher 的精确检验三种方法中，似然比卡方最为准确，即使在小样本中也是如此。而 Fisher 的确切概率法的分布假设在很多时候并不成立，因此建议大家多参考似然比卡方的结果。

第二节　配对四格表卡方检验

一、方法原理

又称为 McNemar 检验，将行变量与纵变量不一致的总例数(b + c)视为固定值，在此条件下来进行推断无需考虑两变量一致的总例数 a 和 d 的大小，这类方法在统计学上称为条件推断方法。常用于配对设计的计数资料，即比较两种检验方法、培养方法或诊断方法，其资料特点为对样本中各观察单位分别用两种方法处理，然后观察两种处理方法的计数结果。

二、分析示例

设有 132 份食品标本，把这份标本一分为二，分别用两种检验方法作沙门菌检验，检验结果如下表所示，试比较两种检验方法的阳性结果是否有差别？(《卫生统计学》第六版例 9-6，P165)

两种检验方法检验结果比较

甲法(a)	乙法(b)		合计
	阳性(b=1)	阴性(b=0)	
阳性(a=1)	80	10	90
阴性(a=0)	31	11	42
合计	111	21	132

三、研究假设

研究问题 1：两种方法检验的阳性结果是否一致(一致性检验)。
研究问题 2：两种方法检验结果的一致性的程度如何(Kappa 检验)。

四、数据录入

1. 变量视图

名称 a　　标签 甲法 值 0 = 阴性 1 = 阳性
名称 b　　标签 乙法 值 0 = 阴性 1 = 阳性
名称 weight　　标签 权重

	名称	类型	宽度	小数	标签	值	缺
1	a	数值(N)	8	0	甲法	{0, 阴性}...	无
2	b	数值(N)	8	0	乙法	{0, 阴性}...	无
3	weight	数值(N)	8	0	权重	无	无

2. 数据视图

	a	b	weight
1	1	1	80
2	0	1	31
3	1	0	10
4	0	0	11

注意数据视图与四格表的差别。

五、操作流程

数据 — 加权个案(W)

频率变量(F):权重weight — 确定

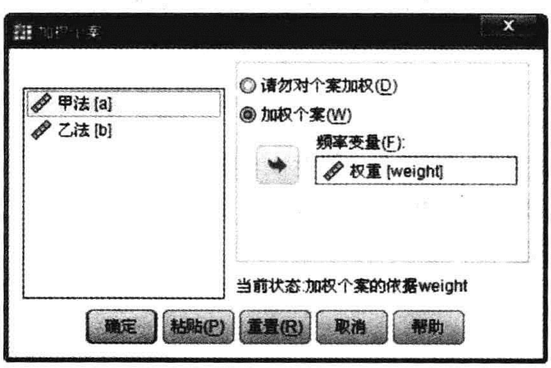
分析 — 描述统计 — 交叉表(C)

行(S):甲法a　　列(C):乙法b

统计量(S) —☑Kappa(K) ☑McNemar(M)— 继续

确定

1. 下图是对数据的预定义,在数据(D)下拉菜单框中完成,即设置权重变量,该操作是对四格表运算的基本操作。



2. 下图是四格表检验的基本对话框,主要设置行变量和列变量,在配对四格表中,行变量和列变量的地位是相等的,本例中甲法和乙法的地位是相等的,因此无需特别区别,选入即可。行变量和列变量的设置只对结果的排列产生影响,并不影响其统计结果。



3. 下图是选择相应统计量。



Kappa(K)检验:又称一致性检验,研究两种方法检验的一致性,希望所有的频数都出现在主对角线上,这样一致性最好。一般认为 $Kappa \geq 0.75$ 时表示两者一致性较好,$0.4 \leq Kappa < 0.7$ 时一致性一般,$Kappa < 0.4$ 时表示两者一致性较差。

McNemar(M)检验:配对卡方检验,研究两者在哪些地方不一致,即两者的诊断结果有怎样的偏向,计算时只利用不在主对角线的数据。

六、结果解释

1. 下表是处理记录缺失值情况报告,可见 132 例均为有效值。

案例处理摘要

	案例					
	有效		缺失		合计	
	N	百分比	N	百分比	N	百分比
甲法 * 乙法	132	100.0%	0	.0%	132	100.0%

2. 下表为原始的四格表。

甲法 * 乙法 交叉制表

		乙法		合计
		阴性	阳性	
甲法	阴性	11	31	42
	阳性	10	80	90
合计		21	111	132

3. 配对卡方检验结果 $P=0.001<0.01$，认为甲法和乙法两者检验的阳性结果有差别。鉴于甲法阳性率 90/132 =68.20%，乙法阳性率 111/132 =84.09%，可以认为乙法阳性率高于甲法阳性率。

卡方检验

	值	精确 Sig.（双侧）
McNemar 检验		.001[a]
有效案例中的 N	132	

a. 使用的二项式分布

4. Kappa $=0.174<0.4$，说明两者的一致性较差；$P=0.027<0.05$，说明 Kappa 值具有统计学意义。

对称度量

	值	渐进标准误差[a]	近似值 T[b]	近似值 Sig.
一致性度量 Kappa	.174	.086	2.206	.027
有效案例中的 N	132			

a. 不假定零假设

b. 使用渐进标准误差假定零假设

七、注意事项

1. 有时可能出现 Kappa 值与 McNemar 检验结果相矛盾的情况,如 Kappa≥0.75,说明两者一致性较好,但是 McNemar 检验 $P<0.05$,说明两者的检验结果有差别。出现这样矛盾的结果是由于两者对于信息的利用不一致造成的,Kappa 检验会利用列联表中的所用信息,而 McNemar 检验只会利用非对角线单元格上的信息,即它只关心两者不一致的评价信息,因此当两者出现矛盾时,主要参考 Kappa 值。

2. 甲乙两种方法的配对检验中,若其中甲法为金标准(Golden Standard),研究乙法与甲法的一致性情况,并研究是否可以用乙法替代甲法。这里就会出现假阳性和假阴性的概念。

(1)假阳性率(false positive rate, FPR),又称误诊率或第Ⅰ类错误,指实际无病,但被认为有病的百分比。

(2)假阴性率(false negative rate, FNR),又称漏诊率或第Ⅱ类错误,指实际有病,但被认为无病的百分比。

如本例以甲法为金标准,其中 90 个病例当中,乙法将 10 个病例视为非病例,即假阴性率为 10/90 = 11.1%;42 个非病例当中,31 个为被视为阳性,即假阳性率为 31/42 = 73.8%,说明乙法假阳性率较高,容易将没病的人错判为有病。

>>第十章　列联表分析

一、基本概念

列联表(contingency table)是指对一组观察对象，分别观察其两种分类变量的表现，归纳成双向交叉排列的统计表，这类统计表用的描述行变量和列变量之间的关系称列联表，又称交叉表。R×C 联列表可以看成是四格表的一个扩展。

二、统计分析策略

变量的统计性质及其专业属性	列联表分类	可选用的统计方法
X,Y 皆为分类变量且属性不同	双向无序表	卡方检验,Fisher 精确检验
X 为分类变量,Y 为有序变量	单向有序表	秩和检验,Ridit 分析,有序变量的 logistic 回归
X,Y 皆为等级变量且属性不同	双向有序表	关心组间差别,按单向有序的列联表处理
		是否有相关,用 Spearman 秩相关或典型相关分析
		是否有直线变化,用线性趋势检验
X,Y 皆为等级变量且属性相同	双向有序表	一致性检验(即 Kappa 检验)

注:几乎所有的列联表资料都可以用对数线性模型或 logistic 回归模型来分析

第一节　双向无序的列联表分析

一、方法原理

双向无序列联表是指行变量和列变量均为无序分类变量，统计方法和一般四格表卡方检验一致。

二、分析示例

某省随机抽查了 1043 位居民的 ABO 血型和 MN 血型，资料如下表所示。问两种血型的构成比有无差别(或两种血型有无关联性)?

某地 1043 位居民的血型

ABO 血型	MN 血型			
	M	N	MN	合计
O	85	100	150	335
A	56	78	120	254
B	98	132	170	400
AB	23	25	6	54
合计	262	335	446	1043

三、研究假设

H_0:两种血型系统分布无差别。

H_1:两种血型系统分布有差别。

$\alpha = 0.05$

四、数据录入

1. 变量视图

名称 x1 标签 ABO 血型

名称 x2 标签 MN 血型

名称 weight 标签 权重

	名称	类型	宽度	小数	标签	值
1	x1	数值(N)	8	0	ABO血型	无
2	x2	数值(N)	8	0	MN血型	无
3	weight	数值(N)	8	0	权重	无

2. 数据视图

	x1	x2	weight
1	1	1	85
2	2	1	56
3	3	1	98
4	4	1	23
5	1	2	100
6	2	2	78
7	3	2	132
8	4	2	25
9	1	3	150
10	2	3	120
11	3	3	170
12	4	3	6

注意该数据视图和 R × C 表的差别

五、操作流程

数据—加权个案(W)

频率变量（F）: 权重weight — 确定

分析 — 描述统计 — 交叉表(C)

行（S）:ABO血型x1　　列(C)：MN血型x2

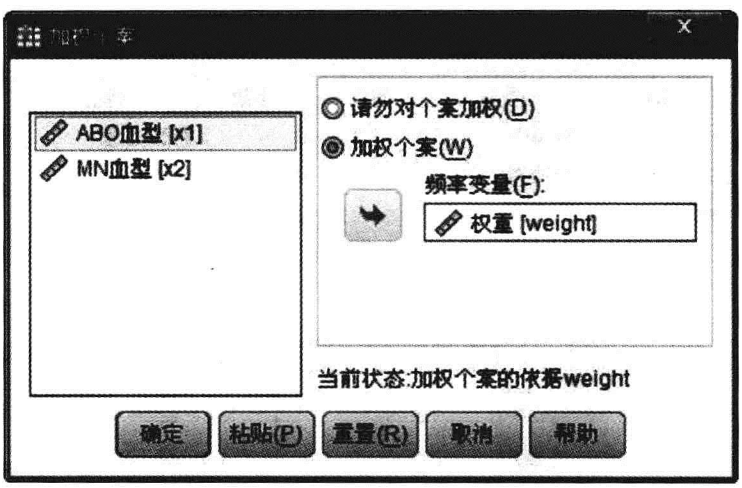
统计量（S）—☑卡方(H)—继续

单元格（E）— 计数 ☑观察值

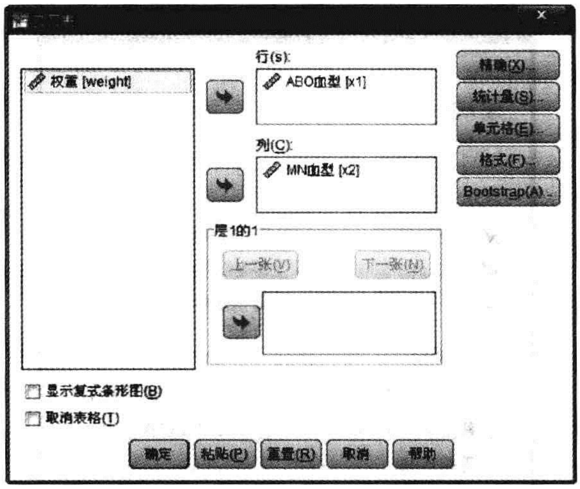
百分比(C) ☑行(R) ☑列(C) ☑总计 — 继续

确定

1. 下图是对数据的预定义，在数据（D）下拉菜单框中完成，即设置权重变量，该操作是对 R × C 表运算的基本操作。



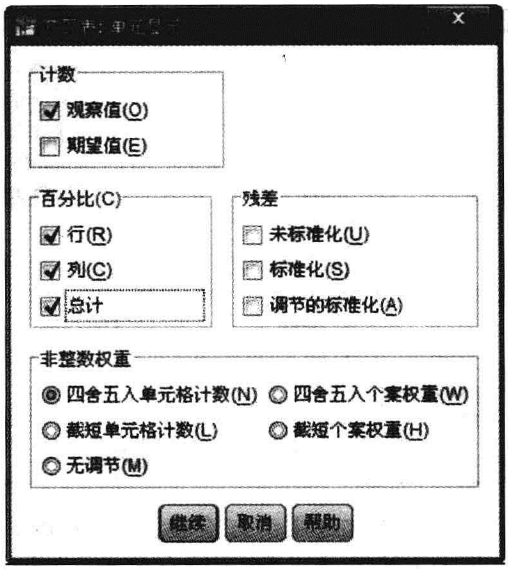
2. 下图是 R × C 表的基本对话框，主要设置行变量和列变量，一般行变量指分组变量，列变量指结果变量，该例中行变量与列变量可以随便制订，行变量和列变量的设置只对结果的排列产生影响，并不影响其统计结果，因此大家可以任意指定行变量和列变量。



3. 下图是对话框选用统计量，本次选用卡方检验(Chi-square)，该检验用于判断行、列变量是否独立，如果数据不满足卡方检验的要求(不能有单元格的期望值 <1，不能有 20% 以上单元格的期望值 <5)，系统会在分析结果中给出提示。



4. 下图中各项是列联表单元格中的描述性指标，**观察值**(observed)指原始数据，**期望值**(expected)指统计过程中的理论数，一般用处不大。各种百分比，如行百分比(Row)、列百分比(Column)及合计百分比(Total)这是 R×C 表中重要的统计描述内容。



六、结果解释

1. 下表是处理记录缺失值情况报告,可见1043例均为有效值。

案例处理摘要

	案例					
	有效		缺失		合计	
	N	百分比	N	百分比	N	百分比
ABO 血型 * MN 血型	1043	100.0%	0	.0%	1043	100.0%

2. 下表标明了各百分比结果,百分数的含义关键是确定分子和分母,分子好确定,也就是各单元格的观察值(计数),分母则需要确定所对应的100%项,主要有三类100%,各横向总计100%,各纵向合计100%,总数共计100%。大家根据需要进行选取,本例当中各百分比并不重要,不需要用百分比作统计描述。

ABO 血型 * MN 血型 交叉制表

			MN 血型			合计
			1	2	3	
ABO 血型	1	计数	85	100	150	335
		ABO 血型 中的 %	25.4%	29.9%	44.8%	100.0%
		MN 血型 中的 %	32.4%	29.9%	33.6%	32.1%
		总数的 %	8.1%	9.6%	14.4%	32.1%

续表

		MN 血型			合计
		1	2	3	
2	计数	56	78	120	254
	ABO 血型 中的 %	22.0%	30.7%	47.2%	100.0%
	MN 血型 中的 %	21.4%	23.3%	26.9%	24.4%
	总数的 %	5.4%	7.5%	11.5%	24.4%
3	计数	98	132	170	400
	ABO 血型 中的 %	24.5%	33.0%	42.5%	100.0%
	MN 血型 中的 %	37.4%	39.4%	38.1%	38.4%
	总数的 %	9.4%	12.7%	16.3%	38.4%
4	计数	23	25	6	54
	ABO 血型 中的 %	42.6%	46.3%	11.1%	100.0%
	MN 血型 中的 %	8.8%	7.5%	1.3%	5.2%
	总数的 %	2.2%	2.4%	.6%	5.2%
合计	计数	262	335	446	1043
	ABO 血型 中的 %	25.1%	32.1%	42.8%	100.0%
	MN 血型 中的 %	100.0%	100.0%	100.0%	100.0%
	总数的 %	25.1%	32.1%	42.8%	100.0%

3. R×C 表卡方检验一般看首行 Pearson 卡方值,但是若 1/5 以上的格子理论频数 <5,或有一个理论频数 <1,则可采用似然比卡方的结果。本例当中 0 单元格的期望计数(理论频数) <5,最小期望计数(最小理论频数)为 13.56,因此采用 Pearson 卡方值 29.925, $P<0.01$,说明两种血型系统分布不同。

卡方检验

	值	df	渐进 Sig.（双侧）
Pearson 卡方	25.925[a]	6	.000
似然比	29.743	6	.000
线性和线性组合	5.533	1	.019
有效案例中的 N	1043		

a. 0 单元格(.0%) 的期望计数 <5,最小期望计数为 13.56

七、注意事项

1. 卡方检验的两两比较：当总体出现差别时，可以进行消耗函数的校正（即对 a 系数进行校正），然后进行多组之间的两两比较。

2. 卡方的精确分割：当 R×C 表资料中含有较多小格频数时，最后结合专业知识对资料进行合并，即对原先分组过细的列联表的列或行进行合并，使调整后的列联表能更好地揭示两个变量之间的关系，合并的原则是：应当将相邻的无显著性差别的列或行进行合并，且这种合并在专业上有实际意义；各分表的自由度之和应等于总表的自由度，各分表的卡方值之和应当等于总表的卡方值。

第二节　单向有序的列联表分析

一、方法原理

单向有序列联表通常指列变量为有序变量的 R×C 表，常见于疗效分析的临床治疗当中，这种资料大家常误用普通卡方检验，其实由于列变量为等级变量，应当采用多组独立样本的秩和检验，属于非参数检验中的一种。

二、分析示例

某医院用 3 种方法治疗慢性喉炎，结果见下表，问这 3 种方法的疗效是否有差别？（《卫生统计学》第六版例 10-6，P185）

3 种方法治疗慢性喉炎的疗效比较

疗效等级	例数			
	甲法(1)	乙法(2)	丙法(3)	合计
无效(疗效=1)	24	20	20	64
好转(疗效=2)	26	16	22	64
显效(疗效=3)	72	24	14	110
治愈(疗效=4)	186	32	22	240
合计	308	92	78	478

三、研究假设

H_0：3 种方法疗效的总体分布位置相同。

H_1：3 种方法疗效的总体分布位置不全相同。

$\alpha = 0.05$

四、数据录入

1. 变量视图

名称 result　　标签 疗效等级 值 1 = 无效 2 = 好转 3 = 显效 4 = 治愈

名称 treat　　标签 治疗方法 值 1 = 甲法 2 = 乙法 3 = 丙法

名称 weight　　标签 权重

	名称	类型	宽度	小数	标签	值
1	result	数值(N)	8	0	疗效等级	{1, 无效}...
2	treat	数值(N)	8	0	治疗方法	{1, 甲法}...
3	weight	数值(N)	8	0	权重	无

2. 数据视图

	result	treat	weight
1	1	1	24
2	2	1	26
3	3	1	72
4	4	1	186
5	1	2	20
6	2	2	16
7	3	2	24
8	4	2	32
9	1	3	20
10	2	3	22
11	3	3	14
12	4	3	22

五、操作流程

数据—加权个案(W)

频率变量(F):权重weight—确定

分析—非参数检验—独立样本

目标:自动比较不同组间的分布

字段:检验字段(T):疗效等级

组(G):治疗方法

设置:自定义检验(C)

比较不同组间分布

☑ Kruskal-Wallis 单因素 ANOVA(K样本)

运行(R)

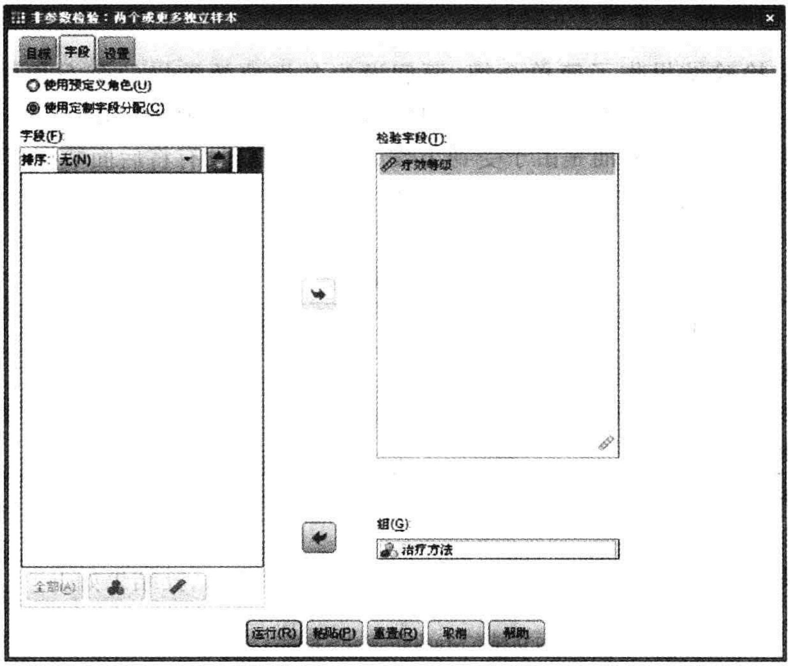
1. 下图为对数据的预定义,在数据(D)下拉菜单框中完成,即设置权重变量,该操作是对 R×C 表运算的基本操作。



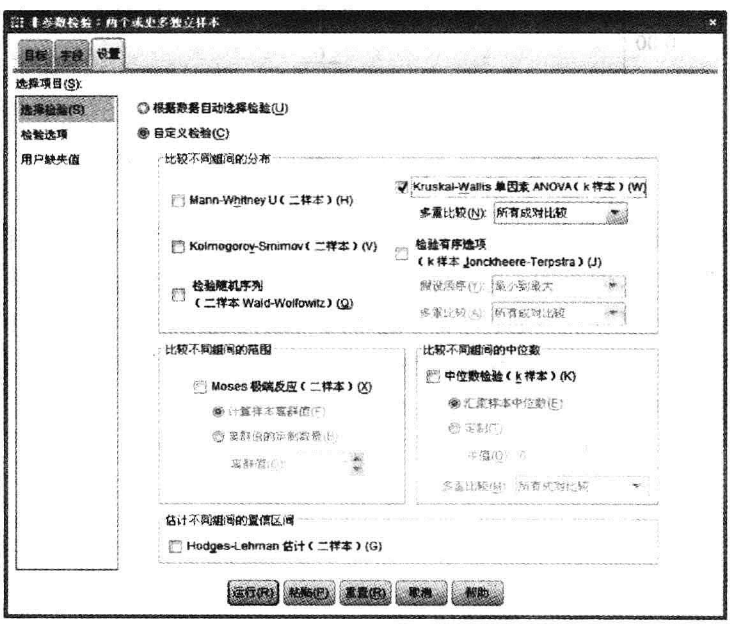
2. 下图指出了非参数检验对于数据的要求:非参数检验不假定您的数据呈正态分布。至于目的,默认为自动比较不同组间的分布,我们先不用管它,因为还需要在其他地方进行设置。



3. 下图为字段选项卡,检验字段(测量指标)框选入疗效等级 result,而组(分组变量)框则选入治疗方法 treat。值得注意的是,有时检验字段的类型必须为数值,组变量的类型必须为字符,变量类型可以在变量视图窗口进行修改。



4. 在选择检验类型中,我们选用 Kruskal-Wallis 单因素 ANOVA,也就是常用的多样本比较的秩和检验。检验方法与多组连续变量但不符合方差分析条件的数据分析方法一致。这种资料经常用于多种药物疗效的比较,但是大家通常选用 R×C 表卡方分析,而非多组独立样本的秩和检验,以致无法得出正确的结论。我们不应选择中位数检验,它虽然直接检验多个样本所代表的各总体的中位数是否相同,但是检验效能很低,很难发现组间的差别。



六、结果解释

1. 下表对检验结果作了简单介绍，原假设为各组疗效等级的分布相同，检验方法为Kruskal-Wallis 检验，即多组独立样本的秩和检验，检验结果为 $P<0.01$（注意：显著性水平显示为0.000时，并非代表0，而是由于受到小数点位数限制，没有将后面数据显示出来），因此可以认为3种方法治疗慢性咽炎的效果有差别。

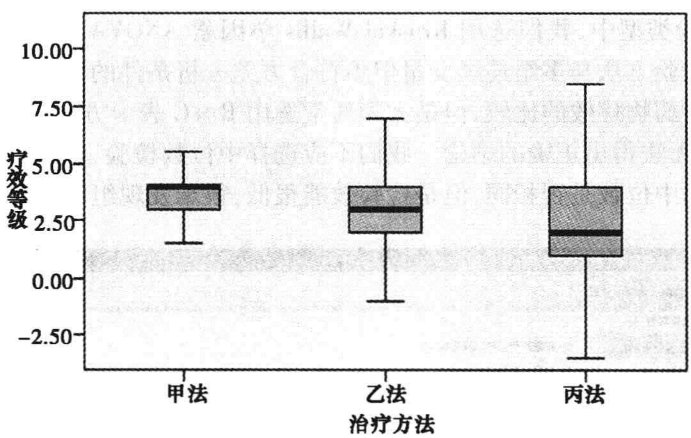
假设检验摘要

	原假设	检验	显著性水平	决策
1	疗效等级的分布在治疗方法的类别间相同	独立样本 Kruskal-Wallis 检验	0.000	拒绝原假设

显示渐进显著性，显著性水平为0.05

2. 在结果窗口双击假设检验摘要表格，弹出详细结果窗口，其结果如下：

独立样本Kruskal-Wallis检验



这是常见的箱式图（box plot），常用于描述不符合正态分布的连续变量的分布特征，此处用于表示多组等级资料。但该图大家需要特别注意，纵轴所代表的疗效等级并非原始数据，而是排序后形成的疗效等级秩次的总体，图形能帮助大家直观地感受各组之间的数据分布。

2. 下表给出了具体的统计值，经过调整的检验统计量为51.388，相当于教科书中的 H 值，$P<0.01$，可以认为3种方法治疗慢性咽炎的效果有差别。

总计 N	478
检验统计	51.388
自由度	2
渐进显著性水平(双侧检验)	0.000

检验统计量已针对结进行了调整

七、注意事项

在多个总体比较中,只能判断它们是否存在差异,若多组比较后总体有差异,接下来的问题应当进行两两比较,来判断到底哪些组别之间有差异,但由于这方面有一定的争议,包括SAS,SPSS在内的所有权威统计软件均未提供该功能,此时可以采取如下策略:

直接使用两组比较的方法进行两组间的非参数检验,但会涉及到控制第一类错误的问题,但是,由于非参数检验方法相对而言检验效能会略低一些,因此对是否一定要调整 α 水准存在争议,一般而言:

(1)如果样本量较小,则不一定需要调整 α 水准,直接比较即可,这样可以补偿非参数检验效能不足所带来的损失。

(2)如果样本量较大,比如每组均在几十例以上,则必须调整 α 水准。

第三节　双向有序且属性不同的列联表分析

一、方法原理

当列联表列变量为有序变量(等级变量),且行变量也为属性不同的有序变量时,我们对这种资料有两种处理方法,第一种是将行变量视为无序变量,则资料转化为单向有序列联表,我们可以采用多组独立样本秩和检验进行分析;第二种我们可以将行变量视为有序变量,来考察这行变量和列变量之间的关系,采用 Spearman 等级相关分析。

二、分析示例

案例 11-3　现有一份 170 例某病患者的治疗效果资料,按年龄和疗效两种属性交叉分类,结果见下表(《卫生统计学》第六版 案例 11-3,P207)。

170 例某病患者的治疗效果资料

患者年龄(岁)	疗效			
	无效(0)	好转(1)	治愈(2)	合计
<18(1)	5	32	20	57
18~(2)	30	38	10	78
50~(3)	15	10	10	35
合计	50	80	40	170

三、数据录入

1. 变量视图

名称 age　标签 患者年龄

名称 result　标签 疗效

名称 weight　标签 权重

	名称	类型	宽度	小数	标签	值
1	age	数值(N)	8	0	患者年龄	无
2	result	数值(N)	8	0	疗效	无
3	weight	数值(N)	8	0	权重	无

2. 数据视图

	age	result	weight
1	1	0	5
2	2	0	30
3	3	0	15
4	1	1	32
5	2	1	38
6	3	1	10
7	1	2	20
8	2	2	10
9	3	2	10

四、数据预处理

数据—加权个案(W)

频率变量(F):权重weight——确定

这是对数据的预定义,在数据(D)下拉菜单框中完成,即设置权重变量,该操作是 R×C 表运算的基本操作。



五、研究方向1:关心组间差别(秩和检验)

(一)研究假设

H_0:3个年龄组疗效的总体分布位置相同。

H_1:3个年龄组疗效的总体分布位置不全相同。

$\alpha = 0.05$

(二)操作流程

分析—非参数检验—独立样本

目标:自动比较不同组间的分布

字段:检验字段(T):疗效

组(G):患者年龄

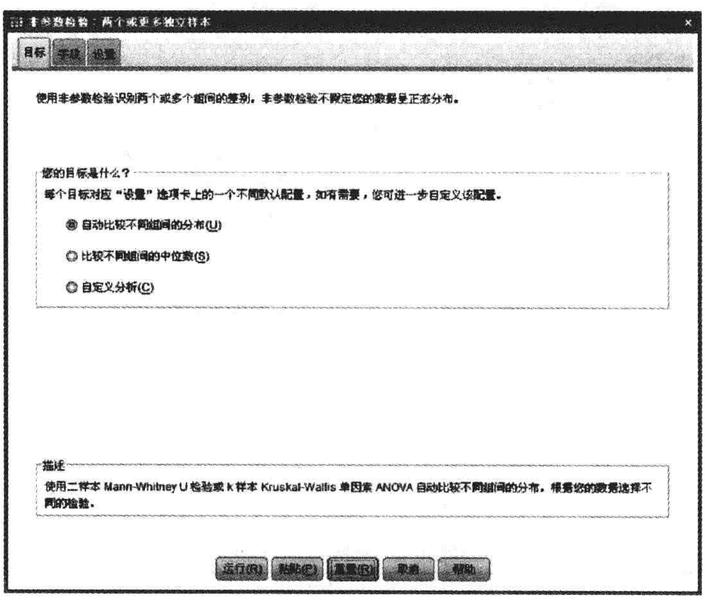
设置:自定义检验(C)

比较不同组间分布

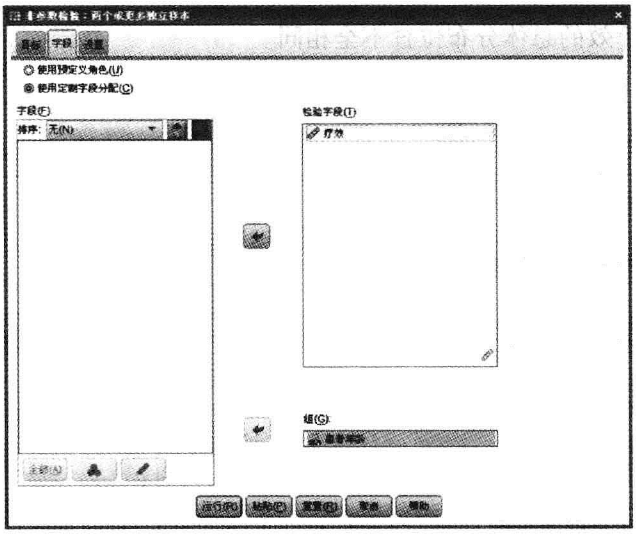
☑ Kruskal-Wallis 单因素 ANOVA(K样本)(W)

运行(R)

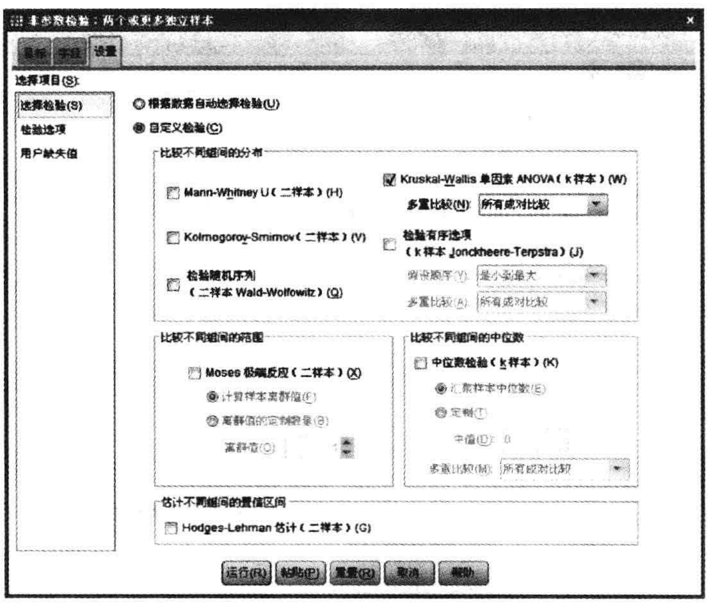
1. 下图指出了非参数检验对于数据的要求:非参数检验不假定您的数据呈正态分布。至于目的,默认为自动比较不同组间的分布,我们先不用管它,因为还需要在其他地方进行设置。



2. 下图为字段选项卡，检验字段（测量指标）框选入疗效 result，而组（分组变量）框则选入患者年龄 age。值得注意的是，有时检验字段的类型必须为数值，组变量的类型必须为字符，变量类型可以在变量视图窗口进行修改。



3. 我们的研究问题为 3 个年龄组之间疗效是否存在差异，因此将 3 个年龄组之间视为无序，该资料转变成单向有序的列联表分析，所采用的方法为多组独立样本的秩和检验，所以检验方法选用 Kruskal-Wallis 单因素 ANOVA。可见同一数据可以根据分析目的的不同可采用不同的方法，从而得出不同的结论。我们不应选择中位数检验，它虽然直接检验多个样本所代表的各总体的中位数是否相同，但是检验效能很低，很难发现组间的差别。



(三)结果解释

1. 下表对检验结果作了简单介绍,原假设为各年龄组的疗效的分布相同,检验方法为Kruskal-Wallis检验,即多组独立样本的秩和检验,检验结果为 $P<0.01$(注意:显著性水平显示为0.000时,并非代表0,而是由于受到小数点位数限制,没有将后面数据显示出来),因此可以认为各年龄组之间的疗效有差别。

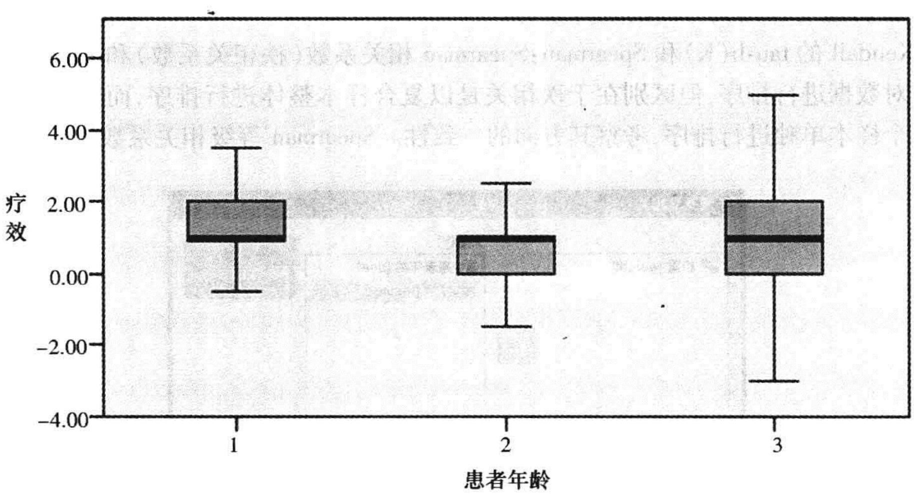
假设检验摘要

	原假设	检验	显著性水平	决策
1	疗效的分布在患者年龄的类别间相同	独立样本 Kruskal-Wallis 检验	0.000	拒绝原假设

显示渐进显著性。显著性水平为0.05

2. 在结果窗口双击假设检验摘要表格,弹出详细结果窗口,其结果如下:

独立样本Kruskal-Wallis检验



这是常见的箱式图(box plot),常用于描述不符合正态分布的连续变量的分布特征,此处用于表示多组等级资料,但该图大家需要特别注意,纵轴所代表的疗效并非原始数据,而是排序后形成的疗效秩次的总体,图形能帮助大家直观地感受各组之间的数据分布。

3. 下表给出了具体的统计值,调整后检验统计量为17.605,相当于教科书中的 H 值,$P<0.01$,可以认为3个年龄组之间的治疗效果有差别。

总计 N	170
检验统计	17.605
自由度	2
渐进显著性水平(双侧检验)	0.000

检验统计量已针对结进行了调整

六、研究方向2:关心两变量之间是否相关(秩相关)

(一)研究假设

研究问题:疗效和年龄之间是否相关

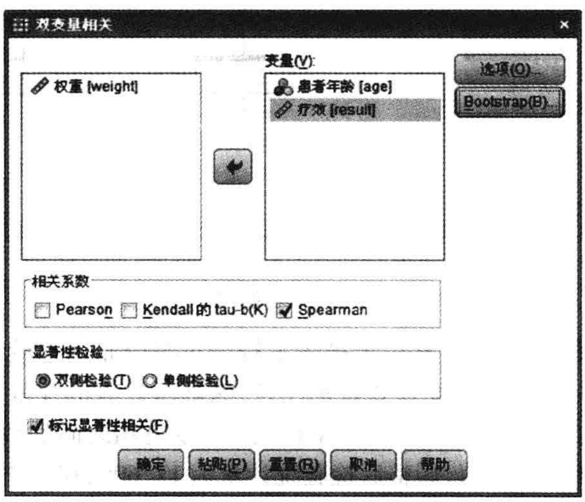
(二)操作流程

分析——相关——双变量
变量(V):患者年龄age　疗效result
相关系数 ☑ Spearman
确定

下图对话框中选择相关性检验方法,这里有相关系数的三个选项,主要体现在对两个变量的要求不同。

(1)Pearson:又称线性相关系数(linear correlation coefficient),是定量表述两个连续变量间线性关系密切程度和相关方向的统计指标。

(2)Kendall 的 tau-b(K)和 Spearman:Spearman 相关系数(秩相关系数)和 Kendall's 等级相关系数均对数据进行排序,但区别在于秩相关是以复合样本整体进行排序,而 Kendall's 等级相关则以每个样本单独进行排序,考察其方向的一致性。Spearman 等级相关系数比较常用。



(三)结果解释

在下面的结果中,可见 Spearman 相关系数为 -0.255,该数值给出了两个信息,相关性的大小为0.255,相关系数绝对值大小为0到1之间,值越大,相关性越强,可见年龄与治疗效果的相关性并不很大;给出了相关的方向,由于 $-0.255<0$,说明两变量呈负相关,即年龄越大,则治疗效果越差。对相关系数的双侧检验的 $P=0.001<0.05$,说明该相关系数具有统计学意义。

相关系数

			患者年龄	疗效
Spearman 的 rho	患者年龄	相关系数	1.000	-.255**
		Sig.(双侧)	.	.001
		N	170	170
	疗效	相关系数	-.255**	1.000
		Sig.(双侧)	.001	.
		N	170	170

**. 在置信度(双侧)为 0.01 时,相关性是显著的

第四节　双向有序且属性相同的列联表分析

一、方法原理

McNemar 检验将行变量与纵变量不一致的总例数视为固定值,在此条件下来进行推断,无需考虑两变量一致的总例数的大小,这类方法在统计学上称为条件推断方法。

Kappa 值实际上为两个差值之比,分子为实际观察到的一致率和可能由偶然机会造成的期望率之差值,差值越大,说明观察到的一致率远比机会造成的期望一致率高,分母为(1 - 期望率),若 Kappa 值较大说明一致性较好。

二、分析示例

对 150 名冠心病患者用两种方法检查室壁收缩运动的情况,检测结果见下表。(《卫生统计学》第六版例 9-7,P167)

两种方法检查室壁收缩运动情况

甲法测定结果	乙法测定结果			
	正常(0)	减弱(1)	异常(2)	合计
正常(0)	60	3	2	65
减弱(1)	0	42	9	51
异常(2)	8	9	17	34
合计	68	54	28	150

三、研究假设

研究问题 1:两种方法检测结果是否一致?

研究问题2:两种方法检测结果的一致性是多大?

四、数据录入

1. 变量视图

名称 x1　　　标签 甲法

名称 x2　　　标签 乙法

名称 weight　　标签 权重

	名称	类型	宽度	小数	标签	值
1	x1	数值(N)	8	0	甲法	无
2	x2	数值(N)	8	0	乙法	无
3	weight	数值(N)	8	0	权重	无

2. 数据视图

	x1	x2	weight
1	0	0	60
2	1	0	0
3	2	0	8
4	0	1	3
5	1	1	42
6	2	1	9
7	0	2	2
8	1	2	9
9	2	2	17

五、操作流程

数据 — 加权个案(W)

频率变量(F):权重weight— 确定

分析 — 描述统计 — 交叉表(C)

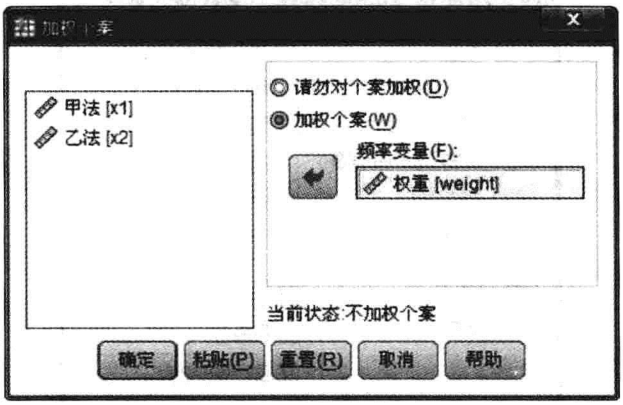
行(S):甲法x1　　列(C):乙法x2

统计量(S)　☑Kappa(K) ☑McNemar(M)—继续

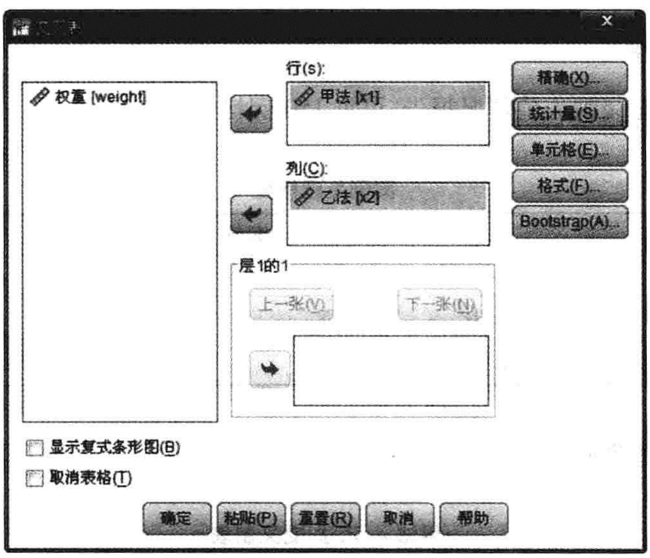
确定

1. 下图是对数据的预定义,在数据(D)下拉菜单框中完成,即设置权重变量,该操作是

R×C表运算的基本操作。



2. 下图是 R×C 表的主对话框，主要设置行变量和列变量，行变量一般指分组变量，列变量指结果变量，本例当中 x1 和 x2 的地位相等，故可以任意设定。行变量和列变量的设置只对结果的排列产生影响，并不影响其统计结果，因此大家可以将行变量与列变量随意调换。



3. 下图是选择统计检验方法，主要有以下两种：

(1)Kappa(K)检验：又称一致性检验，研究两种方法检验的一致性，希望所有的频数都出现在主对角线上，这样一致性最好。一般认为 Kappa≥0.75 时表示两者一致性较好，0.4≤Kappa<0.75 时一致性一般，Kappa<0.4 时表示两者一致性较差。

(2)MaNemar(M)检验：配对卡方检验，研究两者在哪些地方不一致，即两者的诊断结果有怎样的偏向，计算时只利用不在主对角线的数据。



六、结果解释

1. 下表是处理记录缺失值情况报告,可见 150 例均为有效值。

案例处理摘要

	案例					
	有效		缺失		合计	
	N	百分比	N	百分比	N	百分比
甲法 * 乙法	150	100.0%	0	.0%	150	100.0%

2. 下表为原始的列联表。

甲法 * 乙法 交叉制表

计数

		乙法			合计
		0	1	2	
甲法	0	60	3	2	65
	1	0	42	9	51
	2	8	9	17	34
合计		68	54	28	150

3. 配对卡方检验结果 $P=0.086>0.05$,认为甲法和乙法两者检验的结果无差别。

卡方检验

	值	df	渐进 Sig.（双侧）
McNemar-Bowker 检验	6.600	3	.086
有效案例中的 N	150		

4. Kappa =0.676,说明两者的一致性一般;$P<0.01$,说明 Kappa 值具有统计学意义。

对称度量

	值	渐进标准误差[a]	近似值 T[b]	近似值 Sig.
一致性度量　Kappa	.676	.050	11.436	.000
有效案例中的 N	150			

a. 不假定零假设

b. 使用渐进标准误差假定零假设

>>第十一章 Logistic 回归

一、基本思想

多重线性回归模型用于分析一个连续性因变量与一组自变量之间的关系，但是若因变量为分类变量，那么因变量与自变量之间就丧失了这种线性关系，但是经过 Logit 变化，则将模型转变为线性关系，这便是 Logistic 回归模型。

二、医学应用

1. 校正混杂因素：医学研究中，观察对象的某一结局（生存或死亡、阳性或阴性等）常常受到诸多因素的综合影响，包括研究因素与混杂因素。采用 Logistic 回归分析技术，将研究因素、混杂因素及其交互作用均体现在模型中，因此能够在控制混杂因素的作用下，对研究因素与结局变量间的联系作出定量描述。

2. 筛选危险因素：在设计阶段，根据基础理论知识纳入对结局可能有影响的变量，由于一些变量的作用尚不清楚，或纳入变量太多，则需要按照事先规定的检验标准，将有统计意义的变量纳入模型，无统计学意义的变量剔除在外，以保证模型最优。

3. 预测与判断：非条件 Logistic 回归的重要应用之一即是预测与判断，如通过检验，所建立的方程能很好地表达变量间的关系，具有较好的拟合优度。我们给定自变量的数值，则可以通过非条件 Logistic 回归方程计算相应的概率预测值，对个体所属类别作出概率性的判断。但由于条件 Logistic 回归模型不能估计常数项，其结果只能帮助分析变量的效应，不能用于预测。

三、SPSS Logistic 分析模块

1. 非条件 Logistic 回归（二分类 Logistc 回归）：回归模块中的二元 Logistic。

2. 1∶1 条件 Logistic 回归（1∶1 配对 Logistic 回归）：回归模块中多项 Logistic。

第一节 非条件 Logistic 回归

一、方法原理

多重回归分析研究一个正态随机因变量 Y 与一组自变量 X（X_1，X_2，X_3，…，X_P）的数量关系，但我们经常遇到因变量为二分类变量的情况，如发病与否、死亡与否等，需要研究该分类变量与一组自变量之间的关系，则采用二分类 Logistic 回归，又称非条件 Logistc 回归。

二、分析示例

某医师希望研究病人的性别(0 为女性,1 为男性)、心电图检查是否异常(0 为正常、1 为轻度正常、2 为重度正常)、年龄(岁)与患冠心病与否有关,其具体数据见下表。

是否患病 Y	性别 X_1	心电图 X_2	年龄 X_3	是否患病 Y	性别 X_1	心电图 X_2	年龄 X_3
0	0	0	28	0	1	1	32
1	1	0	42	1	0	2	60
0	0	1	46	1	1	2	43
0	1	1	45	0	0	0	59
0	0	0	34	0	1	1	37
1	1	0	44	0	1	0	30
1	0	1	48	1	1	2	47
1	1	1	45	1	0	0	60
0	0	0	38	1	1	1	38
0	1	0	45	0	1	0	34
1	1	1	63	0	0	1	55
0	0	0	51	1	1	1	59
1	1	0	59	0	0	0	50
0	0	2	48	0	1	0	54
0	1	2	35	1	0	1	57
0	0	0	53	1	1	1	60
1	1	0	59	0	0	0	51
1	0	2	57	0	1	0	55
1	1	2	37	1	0	2	46
1	0	0	55	1	1	1	63

三、数据录入

1. 变量视图

名称 y　标签 是否患冠心病　值 0 = 未患病　1 = 患病

名称 x1　标签 性别　值 0 = 女性　1 = 男性

名称 x2　标签 心电图表现　值 0 = 正常　1 = 轻度异常　2 = 重度异常

名称 x3　标签 年龄

	名称	类型	宽度	小数	标签	值	
1	y	数值(N)	1	0	是否患冠心病	{0, 未患病}...	无
2	x1	数值(N)	1	0	性别	{0, 女性}...	无
3	x2	数值(N)	1	0	心电图表现	{0, 正常}...	无
4	x3	数值(N)	2	0	年龄(岁)	无	无

2. 数据视图(部分)

	y	x1	x2	x3
1	0	0	0	28
2	1	1	0	42
3	0	0	1	46
4	0	1	1	45
5	0	0	0	34
6	1	1	0	44
7	1	0	1	48
8	1	1	1	45
9	0	0	0	38
10	0	1	0	45
11	1	1	1	63
12	0	0	0	51
13	1	1	0	59

四、操作流程

分析—回归—二元 Logistic

因变量(D):是否患冠心病 Y

协变量(C): X1 X2 X3

分类(G)

分类协变量(T):X2

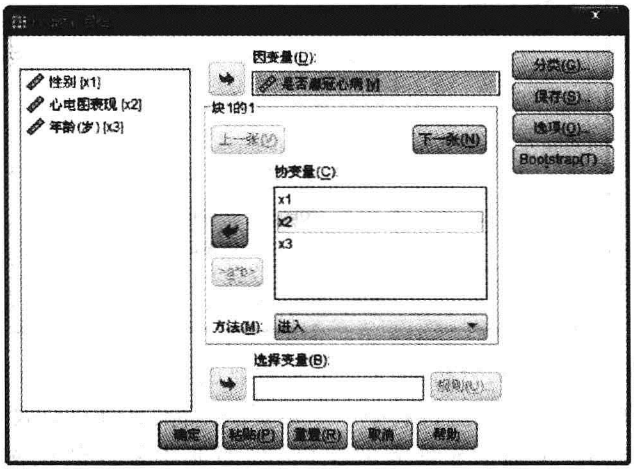
参考类别 ⊙ 第一个 更改(H)—继续

选项 (O) EXP(B)的CI(X) 95%—继续

确定

1. 下图为二分类 Logistic 回归的主对话框,其中因变量框(dependent)中选入二分类因变量,且只能选入一个,本例选入变量是否患有冠心病 y;协变量框(covariates)中选入自变量,本例选入自变量性别 x_1、心电图表现 x_2 和年龄 x_3;方法框(method)用于选入自变量进入模型的方法,一般分为进入法(enter),前进法(forward)和后退法(backward)3 种,前进法和后退法又

可分为条件(conditional)、偏似然比(LR)和 Wald 检验 3 种。如果自变量较少，通常采用进入法；如果自变量太多，则选用前进 LR，前进 LR 相当于多重线性分析中的逐步回归。本例当中采用进入法。



2. 下图是设置哑变量。如果自变量为多分类变量(如血型等)，由于多分类变量和因变量之间不存在线性关系，须用哑变量的方式来分析，系统将产生 K－1 个哑变量(K 为该变量的水平数)。哑变量的设置有对比方式和参考类别两个设置项目，设置方式指各哑变量之间的对比方式，有指示符(indicator)、简单(simple)、差值(difference)、Helmert、重复(repeated)、多项式(polynomial)、偏差(deviation)等形式，我们通常用到的为默认的指示符，并且需要以最后一个或第一个为参考类别，其他水平都与参考水平作比较。本例当中心电图表现 x_2有 3 个水平(正常 $x_2=0$、轻度异常 $x_2=1$、重度异常 $x_2=2$)，按默认的指示符(indicator)方法，若选用最后一个(L)为参考类别，则系统生成的两个哑变量的赋值如下。

1 0　该组反映正常组与重度异常比较。

0 1　该组反映轻度异常与重度异常比较。

0 0　该组为参考的重度异常水平，在统计结果中无法体现。

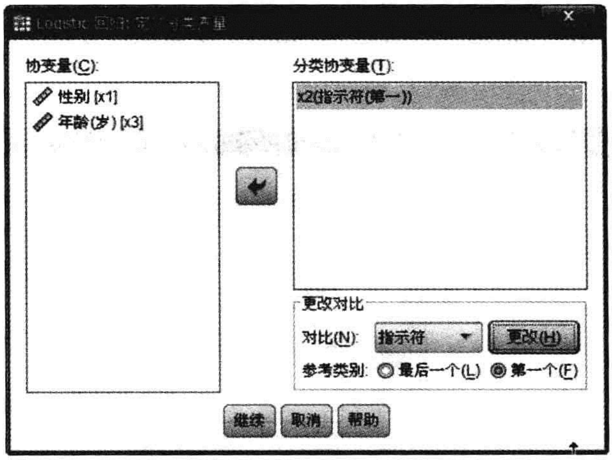
若选用第一个(F)为参考类型，则系统生成的两个哑变量的赋值如下。

0 0　该组为参考的正常水平，在统计结果中无法体现。

1 0　该组反映轻度异常与正常水平的比较。

0 1　该组反映重度异常与正常水平的比较。

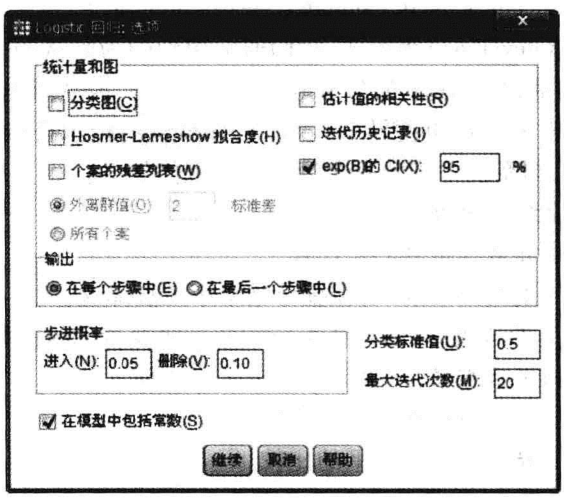
可见，选用不同的参考水平，所代表的含义是不一样的，所以大家在进行结果解释时需要特别注意。



3. 下图将输出 Logistic 回归分析中一个非常重要的指标 OR。Exp(B)的 CI,即 *OR* 值的 95% 可信区间,*OR* 值(比数比)是流行病中一个重要的指标,其计算公式为:

(病例中暴露的比例/病例中非暴露的比例)/(对照中暴露的比例/对照中非暴露的比例)。

如 *OR* =2,则说明病例中暴露于该危险因素的比例为对照中的 2 倍,显示该因素可能与疾病发生有关。



五、结果解释

1. 下表为数据处理情况汇总,模型共 40 例记录纳入分析,0 例缺失。

案例处理汇总

未加权的案例[a]		N	百分比
选定案例	包括在分析中	40	100.0
	缺失案例	0	.0
	总计	40	100.0
未选定的案例		0	.0
总计		40	100.0

a. 如果权重有效,请参见分类表以获得案例总数

2. 下表为因变量赋值的情况,请注意二分类 Logistic(Binary Logistic)过程以因变量较大取值的概率 $P(Y=1)$,而不是以 $P(Y=0)$ 来建立模型,因此在赋值的时候,有必要检查一下结果,确保分析结果的解释正确。大家在建立数据时,将患病用 1 赋值,未患病用 0 来赋值,这样避免麻烦。

因变量编码

初始值	内部值
未患病	0
患病	1

3. 对自变量中分类变量编码进行说明,本例采用指示符(indicator)为编码方法,采用以第一分类为参照(即以最小值 0 为参照)进行编码,形成了两个哑变量,即将变量心电图表现 x_2 变换成两个变量联合进行表示,当心电图为正常($x_2=0$)时,两个哑变量均为 0;当心电图为轻度异常($x_2=1$)时,第一和第二哑变量分别为 1 和 0;当心电图为重度异常($x_2=2$)时,第一和第二哑变量分别为 0 和 1。这样便将默认的连续变量定义为分类变量。

分类变量编码

		频率	参数编码	
			(1)	(2)
心电图表现	正常	19	.000	.000
	轻度异常	13	1.000	.000
	重度异常	8	.000	1.000

4. 块 0:起始块

(1)开始进行模型拟合,即步骤 0,首先给出的模型不含任何自变量,而只有常数项(即无效模型)时的输出结果。该表输出预测分类结果,可见当模型不包含任何自变量时,所有观察对象皆被预测为患病,总的预测准确率为 50.0%。

分类表[a,b]

已观测			已预测		
			是否患冠心病		百分比校正
			未患病	患病	
步骤 0	是否患冠心病	未患病	0	20	.0
		患病	0	20	100.0
	总计百分比				50.0

a. 模型中包括常量

b. 切割值为 0.500

(2)下表给出了该模型中参数的检验结果,此处只有常数项,系数为 0,由于为常数项,有无统计学意义关系不大。

方程中的变量

		B	S.E.	Wals	df	Sig.	Exp (B)
步骤 0	常量	.000	.316	.000	1	1.000	1.000

(3)下表反映的是如果将现有模型外的各个变量纳入模型,则整个模型的拟合优度是否有统计学意义,可以看出若引入性别 x_1,则模型 $\chi^2=1.616$, $P=0.204>0.05$,无统计学意义。若将两个哑变量均引入,则模型 $\chi^2=5.271$, $P=0.072>0.05$,无统计学意义。若将心电图表现 x_2 分拆的的两个哑变量单独引入,仍无统计学意义。而将年龄 x_3 引入,则 $\chi^2=6.810$, $P=0.009<0.05$,有统计学意义。

不在方程中的变量

			得分	df	Sig.
步骤 0	变量	x1	1.616	1	.204
		x2	5.271	2	.072
		x2(1)	1.026	1	.311
		x2(2)	2.500	1	.114
		x3	6.810	1	.009
	总统计量		14.300	4	.006

5. 块 1:方法 = 进入法

(1)步骤 1 表示开始向模型中引入自变量的结果,由于我们采用了进入(enter)法来引入变量,即强迫所有变量同时进入模型。综合检验采用 3 个统计结果:步骤(step)统计量为每一步与前一步相比的似然比检验结果,块(block)统计量指若将块 0 与块 1 相比较的似然比检验结果,而模型(model)统计量则是上一个模型与现在方程中变量有变化后模型的似然比检验

结果。由于本例采用进入法,3 个统计量及假设检验结果完全一致,$\chi^2=18.039$, $P=0.001<0.05$,说明 x_1, x_2, $x_3$3 个变量中至少有一个有统计学意义。

模型系数的综合检验

		卡方	df	Sig.
步骤 1	步骤	18.039	4	.001
	块	18.039	4	.001
	模型	18.039	4	.001

(2)下表为模型汇总,即模型情况简报, -2 倍的似然对数值(-2 Log likelihood)为 37.413, Cox & Snell R^2 和 Nagelkerke R^2 为两个伪决定系数(“伪”,以示与线性回归模型中的决定系数相区别),伪决定系数从不同的角度反映了当前模型中自变量解释了因变量总变异的比例,但对于 Logistic 回归而言,通常看到的模型伪决定系数的大小不像线性回归模型中的决定系数那么大。

模型汇总

步骤	-2 对数似然值	Cox & Snell R 方	Nagelkerke R 方
1	37.413[a]	.363	.484

a. 因为参数估计的更改范围小于 .001,所以估计在迭代次数 6 处终止

(3)下表为现在模型对因变量的分类预测情况,从预测分类表可以看出,预测准确度从块 0(模型只含有常数项)的 50% 上升到 70%,说明新变量的引入对改善模型预测效果有意义。

分类表[a]

已观测			已预测 是否患冠心病 未患病	已预测 是否患冠心病 患病	已预测 百分比校正
步骤 1	是否患冠心病	未患病	14	6	70.0
		患病	6	14	70.0
	总计百分比				70.0

a. 切割值为.500

(4)下表是 Logistic 回归分析中最重要的部分,包括了最终引入模型的变量及常数项的偏回归系数值(β),标准误(SE),Wald 卡方值(Wals),自由度(*df*),*P* 值(Sig.),以及 *OR* 值(Exp(B))。由结果可以看出,变量年龄 x_3 的偏回归系数为 0.163,Wald 检验结果 $P=0.09<0.05$,有统计学意义,*OR* 值为 1.178,*OR* 值 95% 的可信区间为(1.042, 1.330);心电图表现 x_2 中的第二个哑变量也有统计学意义,其偏回归系数为 2.650, $P=0.022<0.05$,*OR* 值为 14.160,*OR*

值的95%可信区间为(1.467,136.709),说明心电图重度异常与正常比较,患心脏病的概率要高。而其他自变量,如性别 x_1,心电图轻度异常与正常比较均没有统计学意义。

方程中的变量

		B	S. E.	Wals	df	Sig.	Exp (B)	EXP(B) 的 95% C. I.	
								下限	上限
步骤 1[a]	x1	1.772	.943	3.531	1	.060	5.885	.926	37.384
	x2			5.429	2	.066			
	x2(1)	1.377	.971	2.012	1	.156	3.962	.591	26.557
	x2(2)	2.650	1.157	5.249	1	.022	14.160	1.467	136.709
	x3	.163	.062	6.890	1	.009	1.178	1.042	1.330
	常量	-9.865	3.549	7.728	1	.005	.000		

a. 在步骤 1 中输入的变量: x1, x2, x3

六、注意事项

1. 二分类(Binary Logistic)过程默认以因变量较大取值的概率 *P* 来建立模型,因此我们对因变量0和1赋值时一定要注意,0代表不发生(即不感兴趣的事件,y=0),1代表发生(即感兴趣事件,y=1)。如我们的血红蛋白是否低下为因变量进行模拟,那么低于某值为发生,代表血红蛋白不正常(y=1),否则为不发生,即血红蛋白正常(y=0)。因变量的赋值决定了模型的解释,大家需要特别注意。

2. 和多重线性回归分析一样,Logisitc 回归分析的自变量也有连续变量、等级变量和分类变量3种,对于连续变量和分类变量(转换为哑变量)比较好处理,比较棘手的是等级变量(也称多分类有序变量),如文化程度可以分为文盲、小学、初中、高中及以上,如本例中心电图表现分为正常、轻度异常和重度异常3种。这样的等级资料可以以连续变量的形式引入模型,但其前提条件是等级分组与 Logit (P)呈线性关系,其效应等比例增加(或降低),如果该前提不能满足,则只能将等级变量当作分类变量,用哑变量来进行分析。

第二节　条件 Logistic 回归

一、基本原理

医学研究中,常采用匹配设计,即为病例组的每一个研究对象匹配一个或几个同样特征的未患病者,作为该病例的对照,这样,除了研究因素外,病例与对照的其他特征相同,从而消除"其他特征"的混杂作用,这样的 Logistic 回归称为条件 Logistic 回归。常用的匹配形式为1:1,即一个病例匹配一个对照。

二、分析示例

Mack 等人(19 世纪 70 年代)欲考察服用雌激素与患子宫内膜癌的关系,对退休居住在社区的妇女进行调查。对照匹配的条件如下:与子宫内膜癌患者的年龄相差不超过一岁,婚姻状况相同,居住在同一社区。除是否服用雌激素外,研究的自变量还包括胆囊病史和是否服用其他非雌激素药物,共计 63 对数据。

ID 对子号	case1 患病	age1 年龄	est1 雌激素	gall1 胆囊病史	nonest1 非雌激素药物	case2 对照	age2 年龄	est2 雌激素	gall2 胆囊病史	nonest2 非雌激素药物
1	1	74	1	0	1	0	75	0	0	0
2	1	67	1	0	1	0	67	0	0	1
3	1	76	1	0	1	0	76	1	0	1
4	1	74	1	0	0	0	70	1	1	1
5	1	69	1	1	1	0	69	1	0	1
6	1	70	1	0	1	0	71	0	0	0
7	1	65	1	1	1	0	65	0	0	0
8	1	68	1	1	1	0	68	0	0	1
9	1	61	0	0	1	0	61	0	0	1
10	1	64	1	0	1	0	65	0	0	0
11	1	68	1	1	1	0	69	1	1	0
12	1	74	1	0	1	0	74	1	0	0
13	1	67	1	1	1	0	68	1	0	1
14	1	62	1	1	1	0	62	0	1	0
15	1	71	1	1	1	0	71	1	0	1
16	1	83	1	0	1	0	82	0	0	0
17	1	70	0	0	1	0	70	0	0	1
18	1	74	1	0	1	0	75	0	0	0
19	1	70	1	0	1	0	70	0	0	0
20	1	66	1	0	1	0	66	1	0	1
21	1	77	1	0	1	0	77	1	1	1
22	1	66	1	0	1	0	67	0	0	1
23	1	71	1	0	0	0	72	0	0	0
24	1	80	1	0	1	0	79	0	0	0
25	1	64	1	0	1	0	64	1	0	1

续表

ID 对子号	case1 患病	age1 年龄	est1 雌激素	gall1 胆囊病史	nonest1 非雌激素药物	case2 对照	age2 年龄	est2 雌激素	gall2 胆囊病史	nonest2 非雌激素药物
26	1	63	1	0	1	0	63	1	0	1
27	1	72	0	1	1	0	72	0	0	0
28	1	57	1	0	1	0	57	1	0	1
29	1	74	0	1	1	0	74	0	0	1
30	1	62	1	0	1	0	62	1	0	1
31	1	73	1	0	1	0	72	1	0	1
32	1	71	1	0	1	0	71	1	0	1
33	1	64	0	0	1	0	65	1	0	1
34	1	63	1	0	1	0	64	0	0	1
35	1	79	1	1	1	0	78	1	1	1
36	1	80	1	0	1	0	81	0	0	1
37	1	82	1	0	1	0	82	0	0	1
38	1	71	1	0	1	0	71	0	0	1
39	1	83	1	0	1	0	83	0	0	1
40	1	61	1	0	1	0	60	0	0	1
41	1	71	1	0	1	0	71	0	0	0
42	1	69	1	0	1	0	69	0	1	1
43	1	77	1	0	1	0	76	1	0	1
44	1	64	1	0	0	0	64	1	0	0
45	1	79	0	1	0	0	82	1	0	1
46	1	72	1	0	1	0	72	1	0	1
47	1	82	1	1	1	0	81	0	0	0
48	1	73	1	0	1	0	74	1	0	1
49	1	69	1	0	1	0	68	0	0	1
50	1	79	1	0	1	0	79	0	0	1
51	1	72	1	0	0	0	71	1	0	1
52	1	72	1	0	1	0	72	1	0	1
53	1	65	1	0	1	0	67	0	0	0
54	1	67	1	0	1	0	66	1	0	1
55	1	64	1	1	1	0	63	0	0	1
56	1	62	1	0	0	0	63	0	0	0

续表

ID 对子号	case1 患病	age1 年龄	est1 雌激素	gall1 胆囊病史	nonest1 非雌激素药物	case2 对照	age2 年龄	est2 雌激素	gall2 胆囊病史	nonest2 非雌激素药物
57	1	83	0	1	1	0	83	0	1	0
58	1	81	1	0	1	0	79	0	0	0
59	1	67	1	0	1	0	66	1	0	1
60	1	73	1	1	1	0	72	1	0	1
61	1	67	1	1	1	0	67	1	1	1
62	1	74	1	0	1	0	75	0	0	1
63	1	68	1	1	1	0	69	1	0	1

三、数据录入

1. 变量视图

名称 id　　标签 对子号
名称 case1　　标签 患病
名称 age1　　标签 病例年龄
名称 est1　　标签 病例是否服用雌激素
名称 gall1　　标签 病例是否有胆囊病史
名称 nonest1　　标签 病例是否服用非雌激素药物
名称 case2　　标签 对照
名称 age2　　标签 对照年龄
名称 est2　　标签 对照是否服用雌激素
名称 gall2　　标签 对照是否有胆囊病史
名称 nonest2　　标签 对照是否服用非雌激素药物

	名称	类型	宽度	小数	标签	值
1	id	数值(N)	8	0	对子号	无
2	case1	数值(N)	8	0	患病	无
3	age1	数值(N)	8	0	病例年龄	无
4	est1	数值(N)	8	0	病例是否服用雌激素	无
5	gall1	数值(N)	8	0	病例是否有胆囊病史	无
6	nonest1	数值(N)	8	0	病例是否服用非雌激素药物	无
7	case2	数值(N)	8	0	对照	无
8	age2	数值(N)	8	0	对照年龄	无
9	est2	数值(N)	8	0	对照是否服用雌激素	无
10	gall2	数值(N)	8	0	对照是否有胆囊病史	无
11	nonest2	数值(N)	8	0	对照是否服用非雌激素药物	无

2. **数据视图(部分)**

	id	case1	age1	est1	gall1	nonest1	case2	age2	est2	gall2	nonest2
1	1	1	74	1	0	1	0	75	0	0	0
2	2	1	67	1	0	1	0	67	0	0	1
3	3	1	76	1	0	1	0	76	1	0	1
4	4	1	74	1	0	0	0	70	1	1	1
5	5	1	69	1	1	1	0	69	1	0	1
6	6	1	70	1	0	1	0	71	0	0	0
7	7	1	65	1	1	1	0	65	0	0	0
8	8	1	68	1	1	1	0	68	0	0	1
9	9	1	61	0	0	1	0	61	0	0	1
10	10	1	64	1	0	1	0	65	0	0	0
11	11	1	68	1	1	1	0	69	1	1	0
12	12	1	74	1	0	1	0	74	1	0	0
13	13	1	67	1	1	1	0	68	1	0	1
14	14	1	62	1	1	1	0	62	0	1	0
15	15	1	71	1	1	1	0	71	1	0	1
16	16	1	83	1	0	1	0	82	0	0	0
17	17	1	70	0	0	1	0	70	0	0	1
18	18	1	74	1	0	1	0	75	0	0	0

四、操作流程

转换(T)——计算变量(C)

目标变量(T):case　数字表达式(E):case1,case2 —— 确定

转换(T)——计算变量(C)

目标变量(T):age　数字表达式(E):age1,age2 —— 确定

转换(T)——计算变量(C)

目标变量(T):est　数字表达式(E):est1,est2 —— 确定

转换(T)——计算变量(C)

目标变量(T):gall　数字表达式(E):gall1,gall2 —— 确定

转换(T)——计算变量(C)

目标变量(T):nonest　数字表达式(E):nonest1,nonest2 —— 确定

分析(A)——回归(R)——多项 Logistic

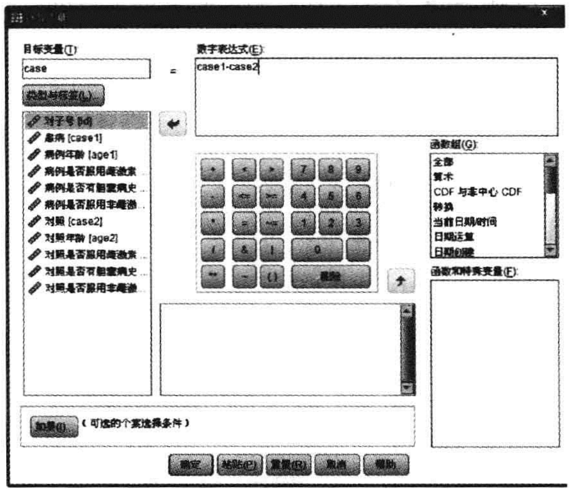
因变量(D):case

协变量(C):age est gall nonest

模型(M) □ 在模型中包含截距(N)—— 继续

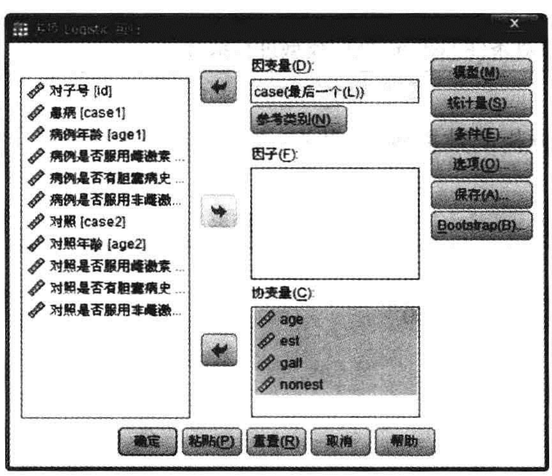
确定

1. 对于 1∶1 配对的 Logistic 回归,我们需要首先对数据进行预处理,生成新的变量以便进行分析,通过转换(T)下拉菜单下的计算变量选项,首先生成新变量 case = case1 - case2。类似,还需要生成 age = age1 - age2; est = est1 - est2; gall = gall1 - gall2; nonest = nonest1 - nonest2。这样就能在数据视图看到新生成的变量 case, age, est, gall, nonest,以下步骤就对这些变量进行条件 Logistic 分析。



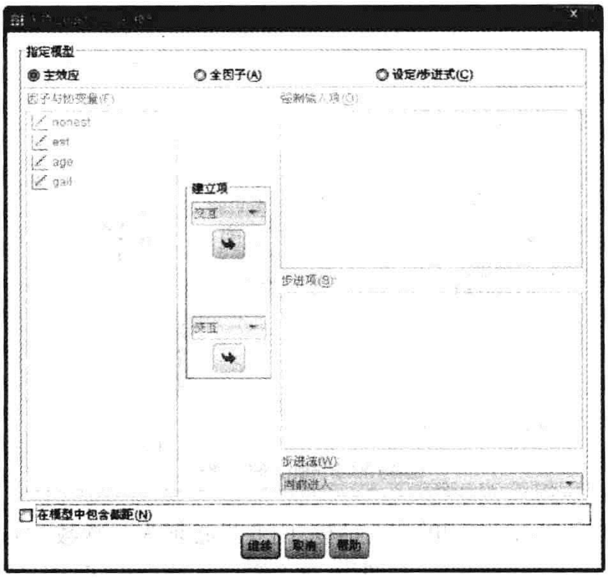
2. 下图是进行条件 Logistic 回归的主对话框,因变量必须为一常数,如本例当中 case = 1,大家需要注意的是 case1 - case2,即病例组 - 对照组,这个顺序不能颠倒了;因子框和协变量框均可选入自变量,但是因子框中的变量系统默认为无序分类资料,会自动将其生成哑变量,而协变量框选入连续变量或有序变量,在条件 Logistic 模型中,分析的自变量均为病例组 - 对照组后新生成的变量,属于连续变量或有序变量,因此我们将 age, est, gall, nonest 均选入协变量框中。

SPSS 中没有专门进行条件 Logistic 分析的过程,但是软件可以调用多项 Logistic 回归过程进行分析,当软件发现因变量为一常数项时,系统自动切换为拟合条件 Logistic 模型,从而通过拟合不包括常数项的模型实现 1∶1 配对的 Logistic 回归。



3. 下图是条件 1:1 回归模型进一步设定，由于模型要求不包含截距，若选入截距项，结果中则会给出警告：*因变量在具有截距项的模型中只有一个有效值，将不对任何模型进行拟合，可通过移去截距对条件Logistic 回归模型进行拟合。*

这里说明了两层意思，若因变量不止一个值，可以进行多分类 Logistic 回归分析；若因变量只有一个值，则需要去掉截距项，系统进行条件 Logisitc 回归分析。



五、结果解释

1. 下表是系统给出的提示，说明因为因变量只有一个值(1)，系统进行条件 Logistic 回归模型拟合。这里需再次提醒大家，进行条件 Logistic 回归分析的两个必要步骤：由病例组-对照组生成新的变量；去除模型的截距项(常数项)。

警告

因变量只有一个有效值。将对条件 Logistic 回归模型进行拟合。

2. 下表给出了进行分析的数据概况，共 63 条记录，无缺失值。

案例处理摘要

		N	边际百分比
case	1.00	63	100.0%
有效		63	100.0%
缺失		0	
总计		63	
子总体		27	

3. 下表是对模型中的所有偏回归系数是否均为 0 进行似然比检验,其结果为 χ^2 = 34.159,P <0.01,说明偏回归系数不全为 0。

模型拟合信息

模型	模型拟合标准	似然比检验		
	-2 倍对数似然值	卡方	df	显著水平
零	87.337			
最终	53.178	34.159	4	.000

4. 下表给出了三类伪决定系数,伪决定系数反映了模型中所有自变量揭示的因变量总变异的比例,但对于 Logistic 回归模型而言,伪决定系数的大小不像线性回归模型中的决定系数那么大。

伪 R 方

Cox 和 Snell	.419
Nagelkerke	.558
McFadden	.391

5. 下表是从当前模型中分别剔除每一个自变量后拟合新的条件 Logistic 回归模型的负二倍似然对数值,用于考察是否可以从当前模型中剔除该自变量。可以看出年龄 age、是否服用非雌激素药物 nonest 的 P 值均 >0.05,提示可以采取逐步回归法对当前模型中的自变量进行筛选。

似然比检验

效应	模型拟合标准	似然比检验		
	简化后的模型的 -2 倍对数似然值	卡方	df	显著水平
age	53.658	.480	1	.488
est	72.013	18.836	1	.000
gall	58.770	5.592	1	.018
nonest	53.279	.102	1	.750

卡方统计量是最终模型与简化后模型之间在 -2 倍对数似然值中的差值。通过从最终模型中省略效应而形成简化后的模型。零假设就是该效应的所有参数均为 0

6. 下表是条件 Logistic 回归分析中最重要的部分,包括了模型中各变量偏回归系数(B),标准误(SE),Wald 卡方值(Wals),自由度(df),P 值(显著水平),以及 OR 值(Exp(B))。由

结果可以看出,变量是否服用雌激素 est 的偏回归系数为 2.698,Wald 检验结果 $P=0.01<0.05$,有统计学意义,*OR* 值为 14.851,*OR* 值 95% 的可信区间为(2.952, 74.723);是否有胆囊病史 gall 也有统计学意义,其系数为 1.836, $P=0.042<0.05$,*OR* 值为 6.270,*OR* 值的 95% 可信区间为(1.066,36.893)。说明服用雌激素和有胆囊病史是子宫内膜癌的危险因素,而其他各项,如年龄 age,是否服用非雌激素类药物 nonest 均没有统计学意义。

参数估计

case		B	标准误	Wald	df	显著水平	Exp(B)	Exp(B) 的置信区间 95%	
								下限	上限
1.00	age	.277	.403	.473	1	.491	1.320	.599	2.908
	est	2.698	.824	10.712	1	.001	14.851	2.952	74.723
	gall	1.836	.904	4.122	1	.042	6.270	1.066	36.893
	nonest	.256	.807	.100	1	.752	1.291	.265	6.279

第十二章　生存分析

一、基本原理

Logistic 回归分析只考虑了终点事件(terminal event)的出现与否，但恶性肿瘤、慢性病或其他情况随访研究中，有时除了考虑终点出现与否，还需考虑观察对象达到终点所经历的时间长短。生存分析(survival analysis)就是将终点事件的出现与否和达到终点所经历的时间结合起来分析的一类统计分析方法。其主要特点是考虑了每个研究对象出现某一结局所经历的时间长短，同时考虑了事件的观察时间和随访时间。

二、基本概念

1. 终点事件(terminal event)　终点事件指由研究者所规定的生存时间的终点。在生存分析中，终点事件是个非常重要的概念，它的定义应尽可能地清楚明了。需要特别提醒的是，终点事件是由研究目的决定的，并非一定是死亡(如研究灯泡寿命)，而死亡也并非一定是终点事件(如肺癌患者死于其他疾病)。发生了终点事件的数据称为完全数据(complete data)，完全数据提供了准确的生存时间。

2. 截尾(censoring)　也称为失访或删失，截尾指在随访研究中，在规定的观察期内，某些观察对象由于各种原因未观察到终点事件发生，并不知道确切的生存时间。发生截尾的数据称为截尾数据(censored data)或删失数据。由于不能提供准确的生存时间，又称为不完全数据(incomplete data)。产生截尾的原因一般为病人失访、病人的生存期超过了研究的终止期。但是截尾数据的价值在于提供了观察期间的信息，生存时间不会短于观察时间。截尾数据常在其右上角标记“+”，表示真实的生存时间未知，只知道比观察到的生存时间要长。

3. 生存时间(survival time)　可以广泛的定义为从规定的观察起点到某一特定终点事件出现的时间长度。其三要素为观察起点、终点事件和时间度量，观察起点和终点事件根据研究目的决定。随机对照临床试验的观察起点通常是随机化的时间；观察性研究中，观察起点可以是发病时间、第一次确诊时间或接受正规治疗的时间等。生存时间的度量单位可以是年、月、日、小时等，生存时间不服从正态分布，常常呈指数分布、Weibull 分布、对数正态分布、对数 Logistic 分布、Gamma 分布或更为复杂的分布，因此需要特殊的统计方法。

三、基本统计指标

以下几个概念是生存分析中非常重要的基本概念，但是这几个概念晦涩难懂，并且术语与解释之间并不贴切，因此大家需要特别注意，仔细揣摩。

1. 生存概率和死亡概率

(1)生存概率(probability of survival):某时段的生存概率表示某时段开始时存活的个体,在该时段结束时仍存活的可能性。如年生存概率表示年初尚存人口存活一年的可能性。$p=$某年活满一年人数/某年年初人口数。

(2)死亡概率(probability of death):某时段的死亡概率表示某时段开始时存活的个体,在该时段内死亡的可能性。如年死亡概率表示年初尚存人口在今后一年内死亡的可能性。$q=$某年内死亡人数/某年年初人口数。

(3)生存概率与死亡概率之间的关系:$p=1-q$。

2. 生存率和死亡概率函数

(1)生存率(survival rate),又称生存函数(survival function)或累计生存率:常用 $S(t)$ 表示,它表示一个个体生存时间长于 t 的概率。生存率具有如下特点:观察起点 $t=0$ 时的生存率为1;当观察期无穷大时,其生存率为0。$S(t)\approx$生存时间长于 t 的患者数/患者总数

(2)死亡概率函数(failure probability function):常用 $F(t)$ 来表示,它表示一个体从开始观察起到时刻 t 的死亡概率,它是一个随时间上升的函数,当 t 无穷大时,死亡概率函数为1。

(3)生存率和死亡概率函数之间的关系:生存率 $=1-$ 死亡概率函数 $S(t)=1-F(t)$

3. 概率密度和风险率

(1)概率密度(probability density),又称概率密度函数(probability density function)或死亡密度函数:常用 $f(t)$ 表示,它表示一个体死于 $(t, t+\Delta t)$ 小区间内的概率极限。$f(t)\approx t$ 时刻开始的区间内死亡患者数/(患者总数 × 区间宽度)

(2)风险率(hazard rate),又称风险函数(hazard function):常用 $h(t)$ 表示,它表示一个生存到时间 t 的个体死于 $(t, t+\Delta t)$ 小区间内的概率极限。$h(t)\approx t$ 时刻开始的区间内死亡患者数/(生存到 t 的患者数 × 区间宽度)

(3)概率密度和风险率均指个体死于 $(t, t+\Delta t)$ 小区间内的概率极限,但是概率密度指一个个体,而风险率指一个生存到 t 时刻的个体。

4. 中位生存期　又称半数生存期:表示恰好有50%的个体尚存活的时间。中位生存期越长,表示疾病的预后越好,反之,中位生存期越短,预后越差。

四、SPSS 生存分析模块

SPSS 主要在分析下拉菜单项中的生存函数(Survival)中实现,其包含的具体统计过程如下:

1. 寿命表　用于分析分组生存资料,求出不同组段的生存率。或者当样本量较大时(如 $n>50$),可以把资料按不同时间段分成几组,观察不同时间点的生存率。

2. Kaplan-Meier　用于样本含量较小时,不能给出特定时间点的生存率,这样就不用担心每个时间段内只有很少的几个观测,甚至没有观测的尴尬局面。

3. Cox 回归　用于拟合 Cox 比例风险模型,这是生存分析中最重要的一种分析方法,它的出现具有划时代的意义,是多因素分析方法中最为常用的一种。

4. Cox 依时协变量　是 Cox 比例模型的进一步发展。当所研究的危险因素其取值随时间而不断变化,或者其作用强度随时间而不断变化时,Cox 模型的适用条件被违反,此时需要对模型加以修正,就必须用到这个过程。举一个典型的例子,临床试验随访资料中经常碰到某研

究对象从安慰剂组退出，跳转至治疗组的资料就应当用此过程来分析。

第一节　寿命表法

一、方法原理

寿命表的基本思想是将整个观察时间划分为很多小的时间段，对于每个时间段，计算所有活过某时间段起点的病例在该时间段内死亡(出现结局)的概率。

二、分析示例

收集了 374 名某恶性肿瘤患者随访资料，取时间区间均为 1 年，整理结果见下表，试估计各年生存率。(《卫生统计学》第六版例 21 - 1，P401)

确诊后年数	0 ~	1 ~	2 ~	3 ~	4 ~	5 ~	6 ~	7 ~	8 ~	9 ~	10 ~
期内死亡数	90	76	51	25	20	7	4	1	3	2	0
期内删失数	0	0	0	12	5	9	9	3	5	5	47

三、数据录入

1. 变量视图

名称 year　标签 生存时间
名称 status　标签 是否删失
名称 number　标签 例数

	名称	类型	宽度	小数	标签	值
1	year	数值(N)	8	0	生存时间	无
2	status	数值(N)	8	0	是否删失	无
3	number	数值(N)	8	0	例数	无

2. 数据视图(部分)

注意原始表格与数据视图之间的差异，比如确诊后 0 年，期间死亡人数为 90，删失人数为 0，则产生两条记录。第一条记录死亡人数，变量 year 赋值为 0，表示确诊后 0 年；变量 status 赋值为 1，表示出现终点事件，即观察到随访对象出现所规定的结局，如本例中死亡；变量 number 赋值为 90，表示出现死亡的人数为 90。第二条记录截尾值(censored value)，表示失访情况，变量 year 赋值为 0，表确诊后 0 年；变量 status 赋值为 0，表示出现失访，原因可能有多种；变量 number 赋值为 0，表示出现失访的人数为 0。如此对确诊后 1 年、2 年等数据进行录入，共形成 22 条记录。

	year	status	number
1	0	1	90
2	0	0	0
3	1	1	76
4	1	0	0
5	2	1	51
6	2	0	0
7	3	1	25
8	3	0	12
9	4	1	20
10	4	0	5
11	5	1	7
12	5	0	9

四、操作流程

数据 — 加权个案(W)

频率变量（F）: 例数number— 确定

分析 — 生存函数（S）— 寿命表(L)

时间（T）: 生存时间 year

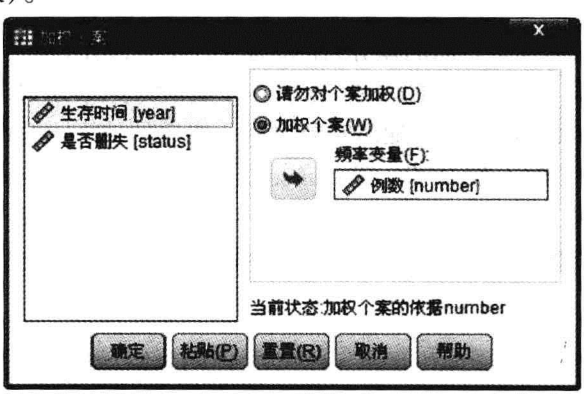
显示时间间隔：0 到 H: 10　步长：1

状态（S）: status(??)

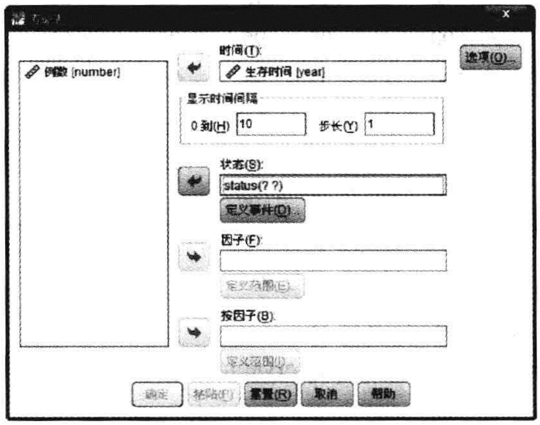
定义事件（D）: 表示事件已发生的值 单值（S）:1 — 继续

确定

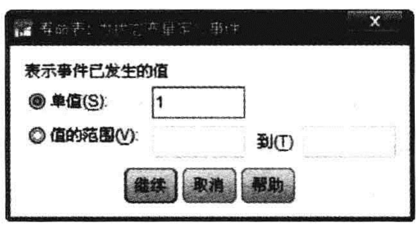
1. 下图是对数据进行预定义，设置权重变量，我们在卡方检验当中经常用到，此处权重变量为变量 number(例数)。



2. 该表为寿命表的主对话框,时间框(time)用于选入生存时间变量 year,如本例中的观察时间;显示时间间隔框(display time intervals)要求选入寿命表中生存时间的范围及步长,本例当中的范围为 0 到 10,步长指观察的间隔,本例当中为 1。如果数据是两年一次收集的,则应当填入 2。状态对话框需要定义终点事件,定义事件需要按定义事件(D)对话框进行进一步定义。



3. 该对话框是进一步定义事件,本例的终点事件为死亡,1 代表死亡,因此在单值(S)填入 1。



五、结果解释

年限表[a]

期初时间		期初记入数	期内退出数	历险数	期间终结数	终结比例	生存比例	期末的累积生存比例	期末的累积生存比例的标准误	概率密度	概率密度的标准误	风险率	风险率的标准误
dimension0	0	374	0	374.000	90	.24	.76	.76	.02	.241	.022	.27	.03
	1	284	0	284.000	76	.27	.73	.56	.03	.203	.021	.31	.04
	2	208	0	208.000	51	.25	.75	.42	.03	.136	.018	.28	.04
	3	157	12	151.000	25	.17	.83	.35	.02	.070	.013	.18	.04
	4	120	5	117.500	20	.17	.83	.29	.02	.060	.013	.19	.04

续表

期初时间	期初记入数	期内退出数	历险数	期间终结数	终结比例	生存比例	期末的累积生存比例	期末的累积生存比例的标准误	概率密度	概率密度的标准误	风险率	风险率的标准误
5	95	9	90.500	7	.08	.92	.27	.02	.022	.008	.08	.03
6	79	9	74.500	4	.05	.95	.25	.02	.014	.007	.06	.03
7	66	3	64.500	1	.02	.98	.25	.02	.004	.004	.02	.02
8	62	5	59.500	3	.05	.95	.24	.02	.013	.007	.05	.03
9	54	5	51.500	2	.04	.96	.23	.02	.009	.006	.04	.03
10	47	47	23.500	0	.00	1.00	.23	.02	.000	.000	.00	.00

a. 中位数生存时间为 2.41

这是寿命表部分,从左到右的各项目为:

(1)期初记入数(number entering this interval):指进入该组段的观察例数,即活到该组段下限的例数。

(2)期内退出数(number withdrawn during interval):指该组段的截尾人数,即失访人数。

(3)历险数(number exposed to risk):指暴露于危险因素的例数,即有效观察人数,等于期初记入数 - 1/2 期内退出数。

(4)期间终结数(number of terminal events):指终点事件的例数,即死亡、复发等研究事件的例数。

(5)终结比例(proportion terminating):指终点事件比例,即各组的死亡概率。

(6)生存比例(proportion surviving):指各组的生存概率,等于 1 - 死亡概率。

(7)期末的累积生存比例(cumul proportion survive at end):指本组段上限的累积生存率,即生存率,由各组的生存概率累计相乘所得。

(8)概率密度(probability density): 常用 f(t)表示,它表示一个体死于(t, t + Δt)小区间内的概率极限。

(9)风险率(hazard rate): 常用 h(t)表示,它表示一个生存到时间 t 的个体死于(t, t + Δt)小区间内的概率极限。

从图上我们可以看出确诊后 5 年,进入该时段的人数为 95 人,失访 9 人,死亡 7 人,因此死亡概率(终结比例)为 0.08,生存概率为 0.92,生存率(期末的累积生存比例)为 0.27,概率密度为 0.022,风险率为 0.08。由此可得到 5 年生存率为 27%,中位数生存时间为 2.41 年。

第二节　单因素生存曲线比较 (Kaplan-Meier 法)

一、方法原理

Kaplan-Meier 法采用乘积极限法(Product-limit estimates)来估计生存率,同时还可以对一个影响因素进行检验。它适用于以个体为单位收集信息的精确生存时间的生存资料,是最为

基本的一种生存分析方法。

二、分析示例

14 例膀胱肿瘤 <3.0 cm 患者和 16 例膀胱肿瘤≥3.0 cm 患者的生存时间(月)如下,试估计两组生存率。(《卫生统计学》第六版. 例 21-2,P403)

肿瘤 <3.0 cm	14	19	26	28	29	32	36	40	42	44^+
	45	53^+	54	59^+						
肿瘤≥3.0 cm	6	7	9	10	11	12	13	20	23	25
	27	30	34	37	43	50				

三、数据录入

1. 变量视图

名称 month 标签 生存时间

名称 status 标签 是否删失 值 0 = 删失 1 = 死亡

名称 group 标签 肿瘤大小 值 1 =" <3.0 cm"　2 = " > =3.0 cm"

	名称	类型	宽度	小数	标签	值	缺失
1	month	数值(N)	8	0	生存时间	无	无
2	status	数值(N)	8	0	是否删失	{0, 删失}...	无
3	group	数值(N)	8	0	肿瘤大小	{1, <3.0cm}...	无

2. 数据视图(部分)

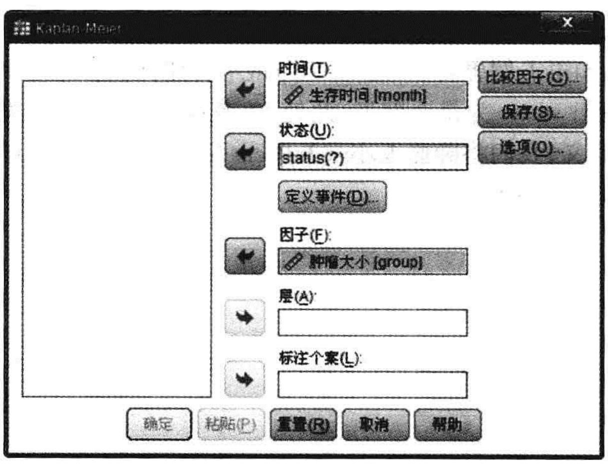
数据有三个变量,如第 1 例患者,肿瘤 <3.0 cm,生存时间为 14 个月,在数据视图窗口输入变量 month 为 14,变量 status 为是否删失,此处 1(表示死亡,即未删失数据);变量 group 对单因素影响因素进行定义,我们将肿瘤大小 <3.0 cm 定义为 1,≥3.0 cm 定义为 2,因此此处赋值为 1。如此共形成 30 条记录。

	month	status	group
1	14	1	1
2	19	1	1
3	26	1	1
4	28	1	1
5	29	1	1
6	32	1	1
7	36	1	1
8	40	1	1
9	42	1	1
10	44	0	1
11	45	1	1
12	53	0	1
13	54	1	1
14	59	0	1
15	6	1	2
16	7	1	2

四、操作流程

分析 — 生存函数（S）— Kaplan-Meier
时间（T）：生存时间 month
状态（U）：status(??)
因子（F）：肿瘤大小 group
定义事件（D）：说明已发生事件的值 单值（S）：1— 继续
比较因子（C）：检验统计量 ☑ 对数秩— 继续
选项（O）：统计量 ☑ 生存分析表☑均值和中位数生存时间
图 ☑ 生存函数— 继续
确定

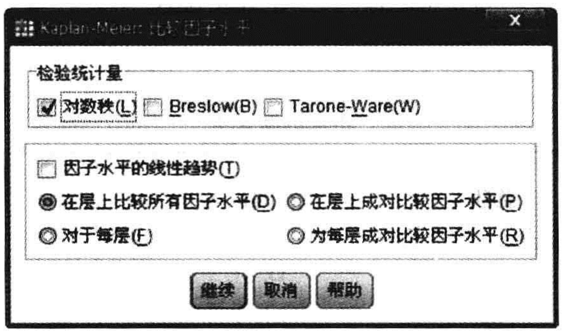
1. 下图为 K-M 检验的主对话框，时间框（survival time）指存活时间，即观察研究的最长时间；状态（status）定义终点事件，此处需要进入定义事件（D）对话框进一步定义；因子框（factor）选入单因素自变量，即定义希望进行比较的研究因素，此处为肿瘤大小 group。



2. 该对话框定义终点事件，此处为说明已发生事件的值，本例定义 1 为死亡，0 即默认为失访。



3. 该对话框用于定义对研究因素(factor)的比较方法,我们通常采用 Log Rank 检验,检验各组生存率分布是否相同,各时间点权重一样。

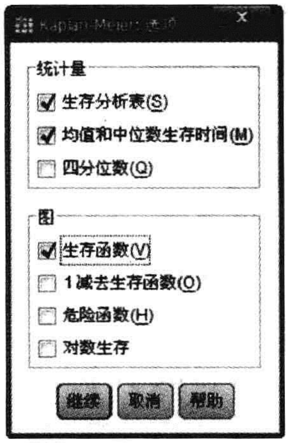


4. 该对话框是对统计量和图形进行定义:

(1)生存分析表(survival table):输出使用乘积极限法(即 Kaplan-Meier 法)计算出生存分析表,类似于寿命表,只不过以每一个个体为单位进行输出,这是区别于以时间段为单位输出的寿命表(life table)的地方。

(2)均数和中位数生存时间(mean and median survival):该选项将给出平均生存时间(该组患者生存率曲线与 X 轴、Y 轴围成的面积)和中位生存时间(该组患者生存时间中位数)及其标准误和可信区间。

(3)生存函数(survival):指累积生存概率曲线。Kaplan-Meier 法估计的是所有死亡时点的生存率,不同死亡时点的生存率逐渐下降,而两个相邻死亡时点之间的生存率都等于前一个较早死亡时点的生存率,所以该生存曲线用水平线连接,呈阶梯形,即每出现 1 例死亡,阶梯就下降一阶。



五、结果解释

1. 下表为原始数据表格的简单归纳，给出了各组的病例总数、终点事件数和删失数（失访数）。

个案处理摘要

肿瘤大小	总数	事件数	删失	
			N	百分比
<3.0 cm	14	11	3	21.4%
≥3.0 cm	16	16	0	.0%
整体	30	27	3	10.0%

2. 下表为生存分析表，可见此处是将病例个体一一给出结果，而不同于寿命表以时间段为单位给出结果。时间（time）指观察时间，状态（status）标明为死亡（终点事件）还是删失（失访），这两项均为原始数据。此时累积的生存比例（cumulative survival）指生存率或生存函数。

生存表

肿瘤大小		时间	状态	此时生存的累积比例		累积事件数	剩余个案数
				估计	标准误		
<3.0 cm	1	14.000	死亡	.929	.069	1	13
	2	19.000	死亡	.857	.094	2	12
	3	26.000	死亡	.786	.110	3	11
	4	28.000	死亡	.714	.121	4	10

续表

肿瘤大小		时间	状态	此时生存的累积比例		累积事件数	剩余个案数
				估计	标准误		
	5	29.000	死亡	.643	.128	5	9
	6	32.000	死亡	.571	.132	6	8
	7	36.000	死亡	.500	.134	7	7
	8	40.000	死亡	.429	.132	8	6
	9	42.000	死亡	.357	.128	9	5
	10	44.000	删失	.	.	9	4
	11	45.000	死亡	.268	.123	10	3
	12	53.000	删失	.	.	10	2
	13	54.000	死亡	.134	.113	11	1
	14	59.000	删失	.	.	11	0
≥3.0 cm	1	6.000	死亡	.938	.061	1	15
	2	7.000	死亡	.875	.083	2	14
	3	9.000	死亡	.813	.098	3	13
	4	10.000	死亡	.750	.108	4	12
	5	11.000	死亡	.688	.116	5	11
	6	12.000	死亡	.625	.121	6	10
	7	13.000	死亡	.563	.124	7	9
	8	20.000	死亡	.500	.125	8	8
	9	23.000	死亡	.438	.124	9	7
	10	25.000	死亡	.375	.121	10	6
	11	27.000	死亡	.313	.116	11	5
	12	30.000	死亡	.250	.108	12	4
	13	34.000	死亡	.188	.098	13	3
	14	37.000	死亡	.125	.083	14	2
	15	43.000	死亡	.063	.061	15	1
	16	50.000	死亡	.000	.000	16	0

3. 下表给出了各组生存时间的均数和中位数，肿瘤<3.0 cm 的生存均值为 38.152 月，生存中位数为 36.000 月；肿瘤≥3.0 cm 的生存均值为 22.313 月，生存中位数为 20.000 月；整体的生存均值为 29.680 月，生存中位数为 28.000 月。但是生存时间一般呈非正态分布，因此采用生存中位数描述资料比较合适，即肿瘤<3.0 cm 组中 50% 的个体存活的时间为 36 月，而肿瘤≥3.0 cm 组中 50% 个体存活的时间为 20 月。

生存表的均值和中位数

肿瘤大小	均值[a]				中位数			
	估计	标准误	95% 置信区间		估计	标准误	95% 置信区间	
			下限	上限			下限	上限
<3.0 cm	38.152	3.740	30.822	45.482	36.000	7.483	21.333	50.667
≥3.0 cm	22.313	3.410	15.628	28.997	20.000	10.000	.400	39.600
整体	29.680	2.860	24.075	35.285	28.000	2.739	22.632	33.368

a. 如果估计值已删失，那么它将限制为最长的生存时间。

4. 下表为两组肿瘤患者整体上的生存曲线比较，结果显示两组肿瘤患者的生存曲线分布差别有统计学意义，Log Rank 法检验统计量$\chi^2=7.369$，$P=0.007<0.01$。

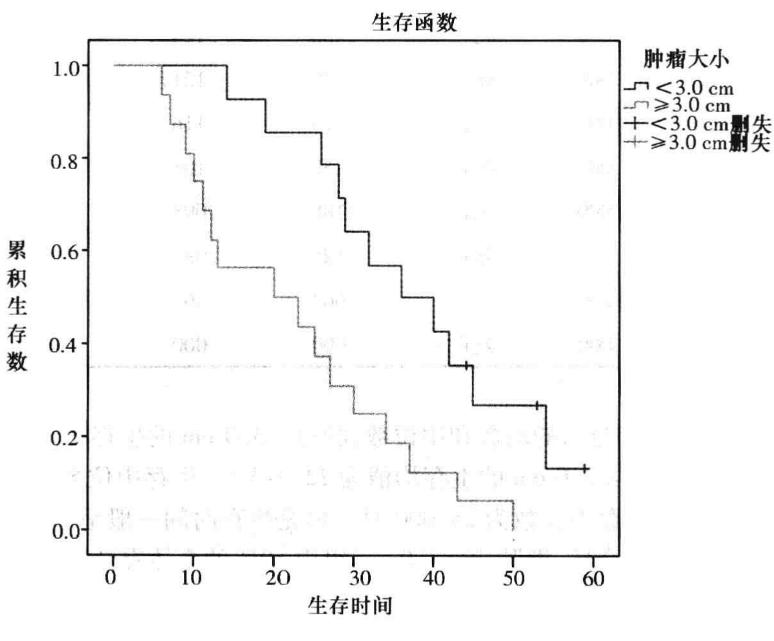
整体比较

	卡方	df	Sig.
Log Rank (Mantel-Cox)	7.369	1	.007

为肿瘤大小的不同水平检验生存分布等同性

Log Rank 检验属于单因素分析方法，应用条件是除比较因素外，影响生存率的各混杂因素组间均衡可比，否则应采用 Cox 比例风险回归模型校正各混杂因素的影响。

5. 下图为两组生存曲线图，配合 Log Rank 检验进行直观表达，可见肿瘤 <3.0 cm 的患者生存状况要明显好于肿瘤≥3.0 cm 的患者。



第三节　多因素生存分析（Cox 回归分析）

一、方法原理

K-M 法只能研究一个因素对生存时间的影响，当对生存时间的影响因素有多个便无能为力，而 Cox 比例风险模型则可以估计多个研究因素对风险率的影响，其过程称为 Cox 回归(Cox Regression)。

二、分析示例

30 例膀胱肿瘤患者的随访记录见表，试进行膀胱肿瘤患者生存情况的影响因素分析。(《卫生统计学》第六版例 21-4，P410)

id	age	grade	size	relapse	start	end	status
1	62	1	0	0	02/10/1996	12/30/2000	0
2	64	1	0	0	03/05/1996	08/12/2000	1
3	52	2	0	1	04/09/1996	12/03/1999	0
4	60	1	0	0	06/06/1996	10/27/2000	0
5	59	2	1	0	07/20/1996	06/21/1998	1
6	59	1	1	1	08/19/1996	09/10/1999	1
7	63	1	1	0	09/16/1996	10/20/2000	1
8	62	1	0	0	09/20/1996	09/18/1999	1
9	50	1	1	0	09/26/1996	03/22/1999	1
10	26	1	1	1	11/04/1996	05/25/2000	1
11	43	2	1	0	01/10/1997	11/08/1999	1
12	62	1	0	0	02/16/1997	11/10/2000	1
13	67	1	0	0	03/09/1997	08/18/2000	1
14	70	2	0	0	03/28/1997	07/20/2000	1
15	56	1	0	1	04/03/1997	11/10/1999	1
16	85	2	0	1	04/15/1997	11/20/1998	1
17	65	1	0	1	08/06/1997	09/28/1999	1
18	54	3	1	1	11/10/1997	12/09/1998	1
19	62	2	0	0	02/19/1998	07/20/2000	1
20	52	3	0	0	03/14/1998	07/02/2000	1
21	63	2	1	0	06/10/1998	09/01/2000	1
22	50	3	1	1	06/15/1998	04/14/1999	1
23	83	2	1	1	09/03/1998	09/20/2000	1
24	61	3	1	0	10/10/1998	06/13/2000	1

续表

id	age	grade	size	relapse	start	end	status
25	57	3	1	1	01/16/1999	12/20/1999	1
26	63	2	0	1	02/17/1999	04/20/2000	1
27	72	3	1	1	05/10/1999	05/12/2000	1
28	56	3	1	1	09/15/1999	06/17/2000	1
29	73	3	1	1	12/19/1999	07/26/2000	1
30	54	3	1	1	03/10/2000	09/20/2000	1

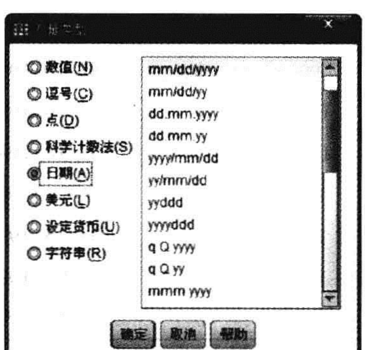
三、数据录入

1. 变量视图

名称 id　　类型 数值
名称 age　　类型 数值
名称 grade　　类型 数值
名称 size　　类型 数值
名称 relapse　　类型 数值
名称 start　　类型 日期
名称 end　　类型 日期
名称 status　　类型 数值

	名称	类型	宽度	小数	标签	值
1	id	数值(N)	8	0		无
2	age	数值(N)	8	0		无
3	grade	数值(N)	8	0		无
4	size	数值(N)	8	0		无
5	relapse	数值(N)	8	0		无
6	start	日期	10	0		无
7	end	日期	10	0		无
8	status	数值(N)	8	0		无

下图是对变量 start 和 end 中对类型进行设定，因为为“日期”类型，该图为日期对话框，可见，日期的输入格式有多种，我们需要选择与原始记录相同的格式，这样有利于输入，本例选中 mm/dd/yyyy，若 1996 年 2 月 10 日，则输入为 02/10/1996。



2. 数据视图(部分)

	id	age	grade	size	relapse	start	end	status
1	1	62	1	0	0	02/10/1996	12/30/2000	0
2	2	64	1	0	0	03/05/1996	08/12/2000	1
3	3	52	2	0	1	04/09/1996	12/03/1999	0
4	4	60	1	0	0	06/06/1996	10/27/2000	0
5	5	59	2	1	0	07/20/1996	06/21/1998	1
6	6	59	1	1	1	08/19/1996	09/10/1999	1
7	7	63	1	1	0	09/16/1996	10/20/2000	1
8	8	62	1	0	0	09/20/1996	09/18/1999	1
9	9	50	1	1	0	09/26/1996	03/22/1999	1
10	10	26	1	1	1	11/04/1996	05/25/2000	1
11	11	43	2	1	0	01/10/1997	11/08/1999	1

四、操作流程

转换 — 日期和时间向导

使用日期和时间进行计算—下一步

计算两个日期之间的时间数—下一步

Date1(D)：　end
减去 Date2(M)：start
单位(U)：　月

结果处理：保留小数位—下一步

结果变量(R)：month
变量标签(L)：生存月数

执行：立即创建变量(C)—完成

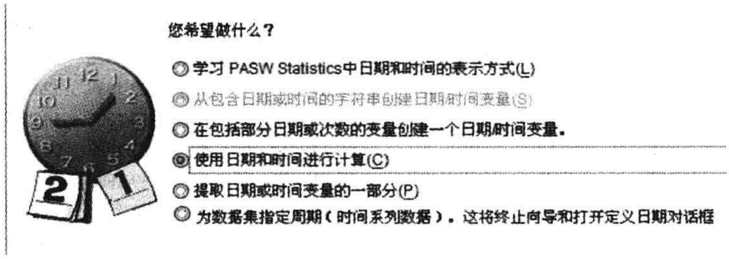
分析—生存函数—Cox 回归
时间(I)：生存月数 month
状态(U)：status(?)
协变量(A)：age grade size relapse

定义事件(F)：表示已发生事件的值 单值(S)：1—继续

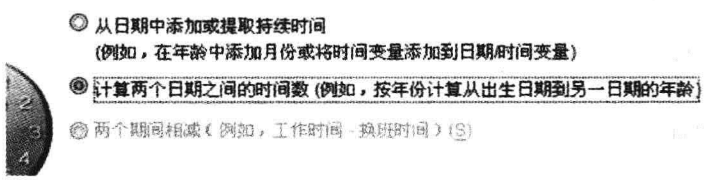
选项(O)　☑ CI 用于 exp(B)95%—继续

确定

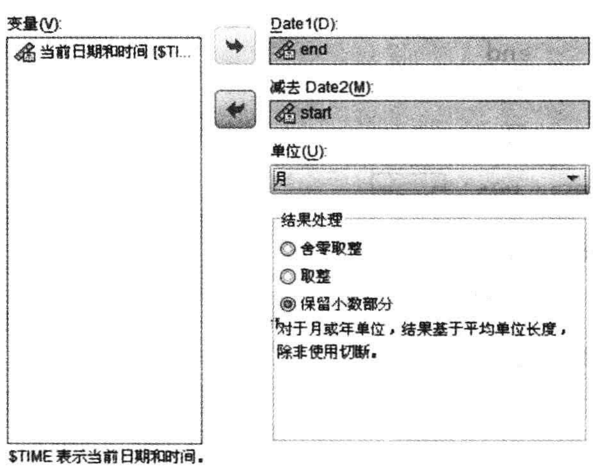
1. 这是对原始数据的预处理，step1 选用使用日期和时间进行计算，因为我们需要获得新变量生存时间，其值等于结束时间-开始时间。



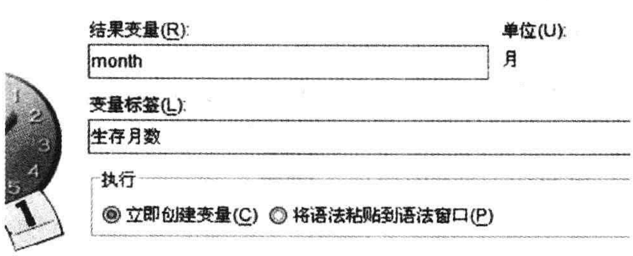
2. 接着进行原始数据的预处理，准备进行计算。



3. 继续进行原始数据的预处理，设定形成的新变量为 end-start，并且单位为月，计算结果保留小数部分，这样会使结果更加精确。

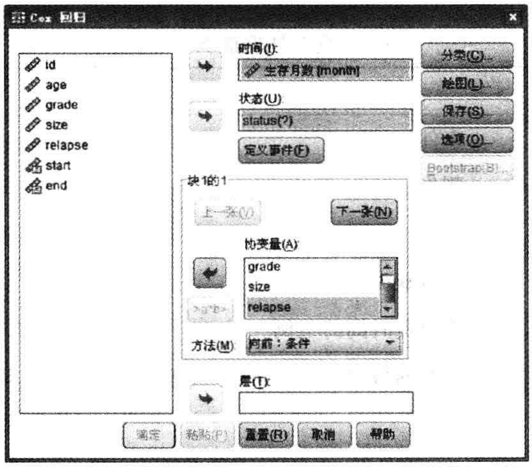


4. 最后对新生成的变量进行定义，其结果变量名称为 month，变量标签为生存月数。



5. 这是 Cox 回归的主对话框，因变量为生存月数 month，也就是刚才预处理后新生成的变

量,同样需要对终点事件进行定义,即定义事件。协变量框选入各类自变量,此处选入变量 age,grade, size, relapse 四个变量。方法(M)选择向前条件法,相当于统计书介绍的逐步回归法。如果自变量较少,也可以采用进入(enter)法,即将协变量框中的所有自变量强行纳入模型。

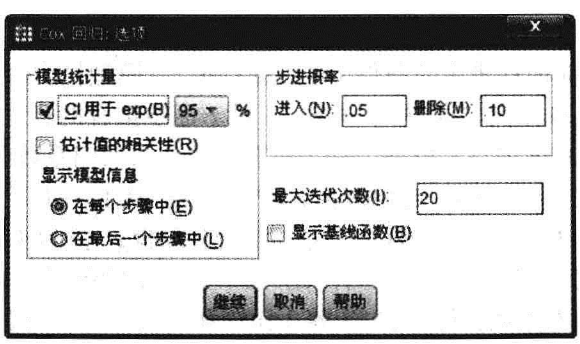


6. 按主对话框中的定义事件(F)进入,即定义研究感兴趣的事件,本例中 1 代表死亡,0 代表失访。



7. 下图选用 exp(B)95%,即 *RR* 值 95% 的可信区间。

RR **值**为相对危险度(relative ratio),适合于队列研究(前瞻性研究),是队列研究中暴露组的发病率与非暴露组的发病率之比。



五、结果解释

1. 下表给出了终点事件为27例，删失为3例，总计30例，并给出了各自百分比。

案例处理摘要

		N	百分比
分析中可用的案例	事件[a]	27	90.0%
	删失	3	10.0%
	合计	30	100.0%
删除的案例	带有缺失值的案例	0	.0%
	带有负时间的案例	0	.0%
	层中的最早事件之前删失的案例	0	.0%
	合计	0	.0%
合计		30	100.0%

a. 因变量：生存月数

2. 块0：起始块

开始进行模型拟合，不过方程只纳入了常数项，因此四个自变量均不在方程中，若将这些方程外的变量分别引入方程中，从而计算出相应的χ^2值和P值，可见grade，size和relapse这三个变量P值均<0.05，可见，在后面的模型拟合中，可以考虑将这三个变量纳入模型当中。

不在方程中的变量[a]

	得分	df	Sig.
age	.402	1	.526
grade	26.094	1	.000
size	7.369	1	.007
relapse	7.349	1	.007

a. 残差卡方 = 37.963，df = 4，Sig = 0.000

3. 块1：方法 = 向前逐步（条件LR）

（1）该模型是按照向前逐步中的条件LR法进行拟合，模型拟合分两步：第一步模型纳入变量grade，第二步纳入size，第三步纳入relapse。对于逐步拟合的过程，我们只需要看最后一步即可，本例则看步骤3，该表是对模型中是否所有的协变量回归系数（常数项除外）全为0进行统计学检验，整体得分为33.981，自由度为3，$P<0.01$。从上一步开始更改是指步骤2纳入了grade和size，而步骤3中引入了relapse，这样统计学检验χ^2值为4.624，$P=0.032<0.05$；从上一块是指从模型仅含有常数项，现在引入了三个自变量的统计学检验，χ^2值为34.421，$P<0.01$。说明步骤3的方程从整体上、与步骤2比较和与块0比较均有统计学意义。

模型系数的综合测试[d]

步骤	-2 倍对数似然值	整体(得分)			从上一步骤开始更改			从上一块开始更改		
		卡方	df	Sig.	卡方	df	Sig.	卡方	df	Sig.
1[a]	120.374	26.094	1	.000	23.161	1	.000	23.161	1	.000
2[b]	113.738	30.782	2	.000	6.636	1	.010	29.797	2	.000
3[c]	109.115	33.981	3	.000	4.624	1	.032	34.421	3	.000

a. 在步骤编号 1：grade 处输入变量

b. 在步骤编号 2：size 处输入变量

c. 在步骤编号 3：relapse 处输入变量

d. 起始块编号 1. 方法 = 向前逐步(条件 LR)

(2)下表是最重要的一个表格,各指标分别为各因子的回归系数估计值(B)、回归系数估计值的标准误(SE)、回归系数估计值的 Wald 检验统计量值(Wald)、自由度(df)、显著性水平即 *P* 值(Sig.)、各因子的相对危险度 *RR* 值(Exp(B))、*RR* 值的 95% 可信区间。

方程中的变量

		B	SE	Wald	df	Sig.	Exp(B)	95.0% CI 用于 Exp(B)	
								下部	上部
步骤 1	grade	1.531	.335	20.947	1	.000	4.625	2.400	8.911
步骤 2	grade	1.623	.366	19.649	1	.000	5.068	2.473	10.386
	size	1.133	.448	6.393	1	.011	3.104	1.290	7.471
步骤 3	grade	1.680	.382	19.385	1	.000	5.367	2.540	11.341
	size	1.078	.460	5.493	1	.019	2.939	1.193	7.242
	relapse	.979	.460	4.525	1	.033	2.662	1.080	6.560

①这里也只看步骤 3,可以得出风险函数的表达式为:

$$h(t) = h_0(t)\exp(1.680 \times grade + 1.078 \times size + 0.979 \times relapse)$$

表达式右边变量的线性组合值越大,则风险函数 h(t)越大,预后越差,故称为预后指数(prognostic index, PI),如 1 号患者 grade = 1,size = 0,replaspe = 0,则预后指数为 PI = 1.680 × 1 + 1.078 × 0 + 0.979 × 0 = 1.680;3 号患者 grade = 2,size = 1,replaspe = 1,则预后指数为 PI = 1.680 × 2 + 1.078 × 0 + 0.979 × 1 = 4.339,可以估计 3 号患者的预后要比 1 号患者差。可按预后指数将观察对象分成若干组,如低危组、中危组和高危组,对制订合理的治疗方案,正确指导患者的治疗,提高生存率有指导意义。

②从 *RR* 值可以看出,若变量肿瘤分级 grade 越严重,则预后越差,大约指数每上升 1 分,则死亡危险是以前的 5.368 倍;若变量肿瘤大小 size 越大,则预后越差,大约指数每上升 1 分,则死亡危险是原来的 2.939 倍;若肿瘤复发,则预后越差,死亡危险是未复发的 2.662 倍。

(3)该表显示各步骤中未纳入方程的自变量检验结果,可见步骤 1 中变量 size 和 relapse 的 P 值分别为0.009 和0.016,考虑将 age 引入方程;而步骤2 中变量 age 的 P 值为0.115,$P>0.05$,不考虑引入模型;而 relapse 的 P 值为0.029,考虑引入模型;步骤3 中 age 的 $P>0.05$,表明模型无需引入变量,模型拟合过程结束。

不在方程中的变量[a,b,c]

		得分	df	Sig.
步骤 1	age	1.061	1	.303
	size	6.844	1	.009
	relapse	5.828	1	.016
步骤 2	age	2.483	1	.115
	relapse	4.769	1	.029
步骤 3	age	3.024	1	.082

a. 残差卡方 = 15.092, df = 3, Sig = 0.002

b. 残差卡方 = 8.518, df = 2, Sig = 0.014

c. 残差卡方 = 3.024, df = 1, Sig = 0.082

(4)下表也是对模型拟合程度的一个检验,如步骤 1,若删除了引入的变量 grade,则 $P<0.01$,说明引入该变量有意义;而步骤 2 中,若删除引入的变量 grade,则 $P<0.01$,若删除引入的变量 size,则 $P=0.010<0.05$,说明引入这两个变量均有意义;步骤 3 中,若分别删除变量 grade、size 和 relapse,P 均 <0.05,说明引入这三个变量均有意义。这只是一个参考表格,不必重视。

如果删除项则建模

项已删除		丢失卡方	df	Sig.
步骤 1	grade	23.161	1	.000
步骤 2	grade	23.276	1	.000
	size	6.682	1	.010
步骤 3	grade	23.985	1	.000
	size	5.767	1	.016
	relapse	4.650	1	.031

(5)下表给出了各自变量的平均值,该表也不太重要。

协变量均值

	均值
age	60.167
grade	1.900
size	.533
relapse	.500

>> 第十三章　诊断试验的统计分析

对疾病的诊断是医学研究中一个非常热门的话题，从临床角度而言，它不仅包括实验室检查，亦包括各种影像诊断，如 X 线诊断、CT 诊断、磁共振成像（MRI）、超声波诊断以及各种放射性核素检查、纤维内镜、电镜等诊断方法。对各种诊断结果的准确性和一致性检验也日显突出，而一般的统计数据则很少涉及这两个方面。本章介绍应用最为广泛的 ROC 曲线和 Bland-Altman 图。

ROC 曲线：对诊断方法的准确性评价，欲知一种检测方法的准确性，则一定需要一个可以信赖的标准（金标准），如同要判定一个考生的考试成绩，就需要知道其考卷的标准答案一样。在临床实践中，金标准可能费时费力，所以试图寻找一个可以替代的检测方法，这就是 ROC 曲线的应用。ROC 曲线要求资料为连续变量或等级变量，以金标准为分组依据，图示灵敏度和特异度来探讨该方法的准确性。同时也可以探讨多种替代方法之间的优劣，所采用的软件为 SPSS。

Bland-Altman 图：很多情况下，需要探讨两种检测方法的一致性，如果两种检测方法的一致性较好，那么则可以相互替代，此时最常用到 Bland-Altman 图，所采用的软件为 GraphPad Prism。

第一节　准确性检验（ROC 曲线）

一、方法原理

（一）ROC 曲线的概括

ROC 曲线是受试者工作特征（receiver operating characteristics）的缩写，它是一种广泛应用的数据分析方法，1950 年应用于雷达信号检测的分析，用于区别“噪声”和“信号”，后来应用于心理学研究。1960 年 Lee Lusted 首先认识到 ROC 分析方法在医学判断疾病方面可能起到的作用。

（二）灵敏度和特异度的基本概念

对于一组经金标准诊断的患者和正常人，如果用一种新的诊断方法试验，其结果可以汇总为下表：

诊断试验结果汇总表

试验	患者	正常人	合计
阳性	a	b	a+b
阴性	c	d	c+d
合计	a+c	b+d	a+b+c+d

真阳性率(灵敏度)=a/(a+c)

真阴性率(特异度)=d/(b+d)

假阳性率(误诊率)=b/(b+d)

假阴性率(漏诊率)=c/(a+c)

(三)ROC 曲线的绘制原理

若检测结果为定量资料或等级资料,当选择不同检测值作为判断阳性、阴性结果的阈值时可以分别计算出相对应的特异度和灵敏度,以(1-特异度)为横轴,灵敏度为纵轴,将坐标轴以(1-特异度,灵敏度)的数据点描绘于平面直角坐标系,将各点连接起来的曲线则为 ROC 曲线。

(四)ROC 曲线的含义

1. 单条 ROC 曲线的解释:一个优良的诊断试验其 ROC 曲线应当从左下角垂直上升至顶线,然后水平方向向右延伸到右上角。如果 ROC 曲线沿着对角线方向分布,表示分类是机遇造成的,正确分类和错分的概率各为 50%,此时该诊断方法完全无效。

2. 两条 ROC 曲线的解释:如果有两种新方法同时测量各标本,则绘制两条 ROC 曲线。如果两条曲线不交叉,那么可以比较两个实验的优劣:更外面的、离对角线更远的曲线,其灵敏度和特异度均高于里面的、离对角线更近的曲线。

二、分析示例

某医师对经过金标准诊断的 55 名患者、45 名正常人分别进行两种诊断实验检查,结果见下表。绘制 ROC 曲线。

100 例受试对象的金标准诊断和两种诊断实验结果

受试对象	诊断结果	检测 1	检测 2	受试对象	诊断结果	检测 1	检测 2
1	患者	112.7	124	9	正常	94.9	120
2	患者	104	135.8	10	患者	140.2	140.9
3	患者	126.7	122.7	11	患者	119.7	142.1
4	患者	123.3	158.4	12	正常	98.6	133
5	患者	120.5	141.2	13	正常	77.3	121.7
6	患者	130.3	131.1	14	患者	139.9	128.8
7	患者	129.6	148	15	正常	97.9	116.6
8	正常	97.9	130.6	16	患者	134.2	130.9

续表

受试对象	诊断结果	检测1	检测2	受试对象	诊断结果	检测1	检测2
17	患者	137.5	150.5	53	患者	115.6	155.4
18	患者	131.2	131	54	患者	110.5	157.4
19	患者	110	140.2	55	患者	127	159.4
20	正常	99.7	117.5	56	患者	131.6	175.7
21	患者	121	135.5	57	患者	128.2	157.2
22	患者	131.1	131.5	58	正常	106.9	141.7
23	患者	108.9	147.5	59	正常	107.9	141
24	患者	121.2	138	60	患者	118.4	153.6
25	正常	83	132.1	61	患者	128	153.9
26	患者	124.3	135.4	62	患者	126.8	154.6
27	正常	102.5	133.9	63	患者	104.9	164
28	正常	104.5	147	64	正常	100.3	129.3
29	患者	128.7	133.8	65	正常	133.4	136
30	患者	130.8	119.3	66	正常	90.6	144.8
31	正常	108.9	108.4	67	正常	102.9	136.6
32	正常	93.2	115.8	68	患者	134.8	165.8
33	正常	101.3	114.7	69	正常	86.4	144
34	患者	138.8	137.1	70	患者	132.8	166.6
35	患者	110.4	141.8	71	正常	107.7	167.5
36	正常	99.8	119.7	72	患者	128.9	144.9
37	正常	108.3	108.7	73	患者	123.1	152.4
38	正常	86	137.9	74	患者	135.7	139.1
39	患者	120.6	125.5	75	患者	124.5	160.6
40	正常	94.9	126.6	76	正常	98.8	142.2
41	患者	102.7	142.8	77	正常	100.2	144.4
42	患者	126.6	147.5	78	正常	105.4	155.4
43	正常	103.2	122.4	79	正常	95.1	155.9
44	患者	123	151	80	患者	110.7	160.9
45	患者	119.9	149.8	81	正常	85.6	149.9
46	患者	95	131.3	82	正常	102.5	132.1
47	患者	143.6	136.2	83	正常	108.9	133.5
48	正常	84	128.3	84	正常	112.2	152.8
49	正常	84.2	138.8	85	正常	102.8	139
50	正常	112.9	126.8	86	患者	119.2	144.6
51	患者	110.5	129.1	87	患者	131.1	154.5
52	患者	126.8	143.4	88	正常	92.4	127.7

续表

受试对象	诊断结果	检测 1	检测 2	受试对象	诊断结果	检测 1	检测 2
89	患者	133.1	157.4	95	正常	96.8	157.4
90	患者	114.6	171.2	96	正常	114.8	142.8
91	正常	94	162.5	97	正常	86.2	144.5
92	患者	131.8	141.9	98	患者	113.1	136.9
93	正常	94.1	142.1	99	正常	88.9	149.8
94	正常	77.4	138.1	100	患者	132.5	158.9

三、数据录入

1. 变量视图

名称 id　标签 受试对象

名称 diag　标签 金标准诊断 值:0 = 正常;1 = 患者

名称 test1　标签 检测 1

名称 test2　标签 检测 2

	名称	类型	宽度	小数	标签	值
1	id	数值(N)	8	0	受试对象	无
2	diag	数值(N)	8	0	金标准诊断	{0, 正常}...
3	test1	数值(N)	8	1	检测1	无
4	test2	数值(N)	8	1	检测2	无

2. 数据视图(部分)

	id	diag	test1	test2
1	1	1	112.7	124.0
2	2	1	104.0	135.8
3	3	1	126.7	122.7
4	4	1	123.3	158.4
5	5	1	120.5	141.2
6	6	1	130.3	131.1
7	7	1	129.6	148.0
8	8	0	97.9	130.6
9	9	0	94.9	120.0
10	10	1	140.2	140.9
11	11	1	119.7	142.1
12	12	0	98.6	133.0
13	13	0	77.3	121.7
14	14	1	139.9	128.8
15	15	0	97.9	116.6

四、操作流程

分析—ROC 曲线图
检验变量(T) ： 检测 1(test1) 检测 2(test2)
状态变量（S）: 金标准诊断 diag
状态变量的值(V)：1
输出
☑ ROC 曲线
☑ 带对角线参考线
☑ 标准误和置信区间（E）
确定

下图为 ROC 曲线的主对话框。

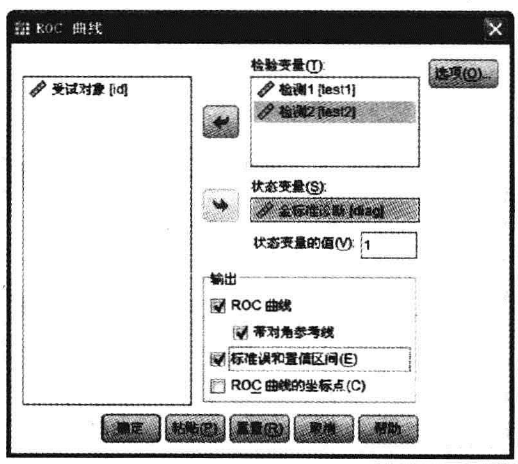
检验变量:即需要研究的检测方法。如本例的两种检测方法:检测 1 和检测 2。

状态变量:通过金标准诊断所确定的各受试对象的结果:患者还是正常人。而下面的状态变量的值(V)需要输入表示患者的值,本例中 1 = 患者,0 = 正常人,所以此处填入 1。

ROC 曲线:为默认选项,即绘制 ROC 曲线图形。

带对角参考线:为 ROC 曲线图形添加对角参考线。

标准误和置信区间:计算和显示曲线下面积、标准误和可信区间。



五、结果解释

1. 下表为资料概括,经过金标准诊断,共有患者 55 例,正常人 45 例。在本例当中检测值越高,就越有可能为患者。

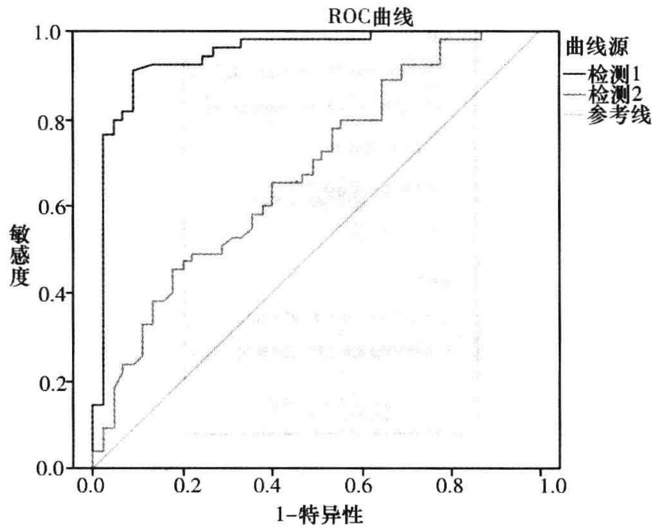
案例处理摘要

金标准诊断		有效的 N(列表方式)
dimension0	正的[a]	55
	负的	45

检验结果变量的值越大,越可以证明实际状态为正

a. 正的实际状态为患者

2. 下图为 ROC 曲线,由此可见检测 1 的效果远远好于检测 2,至于两者面积的具体值,见下表。



结生成的对角段

3. 检测 1 的 ROC 曲线下面积为 0.947,标准误为 0.024,其 95% 的可信区间为(0.900, 0.994);检测 2 的 ROC 曲线下面积为 0.679,标准误为 0.053,其 95% 的可信区间为(0.574, 0.784)。

ROC 曲线下面积取值范围为 0.5 ~ 1.0,一般来说,ROC 曲线下面积在 0.5 ~ 0.7 之间表示诊断价值较低,在 0.7 ~ 0.9 之间表示诊断价值中等,0.9 以上表示诊断价值较高。

检测 1 和检测 2 的渐进 Sig 均 <0.01,该检验的假设是检测方法总体 ROC 曲线下面积是否为 0.5,即该检测方法是否无效,经检验,这两种方法均有效。

曲线下的面积

检验结果变量	面积	标准误[a]	渐进 Sig.[b]	渐近 95% 置信区间	
				下限	上限
检测 1	.947	.024	.000	.900	.994
检测 2	.679	.053	.002	.574	.784

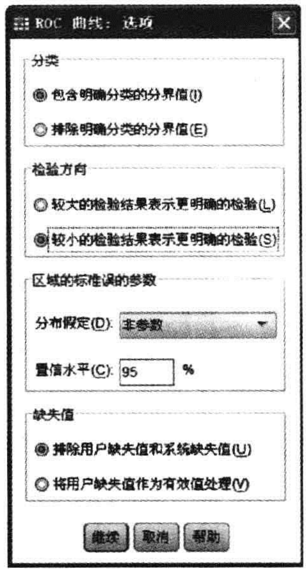
检验结果变量:检测 1, 检测 2 在正的和负的实际状态组之间至少有一个结,统计量可能会出现偏差

a. 在非参数假设下

b. 零假设:实面积 = 0.5

六、注意事项

本例中检测值越高,越可能为患者;还有另外一种情况,检测值越低,越可能是患者,如低血糖患者检测,当血糖值越低,就越可能为患者。此时需要在 ROC 曲线选项对话框中进行选择,见下图:



检测方向

1. 较大的检验结果表示更明确的检验　此为默认选项,表示检测值越大,越可能为患者,如本例。

2. 较小的检验结果表示更明确的检验　如果检测值越小,越可能为患者,则需要选择此项。如低血糖患者。

第二节　一致性检验 (Bland-Altman 图)

本节以 GraphPad Prism 5 为介绍对象,软件的基础知识参阅"第十五章:GraphPad Prism 5 绘图界面介绍"。

一、方法原理

Bland-Altman 分析用于评价测量结果为连续性资料的两种方法一致性,最初由 Bland JM 和 Altman DG 于 1986 年提出,其基本思想是计算两种测量结果的一致性界限(limit of agreement),并用图形的方法直观反映一致性界限,最后结合临床实际,得出两种测量方法是否具有一致性的结论。

二、分析示例

在测量 16 名检查者心功能指标左心室舒张末期容量(EDV)时,分别用多次屏气电影法

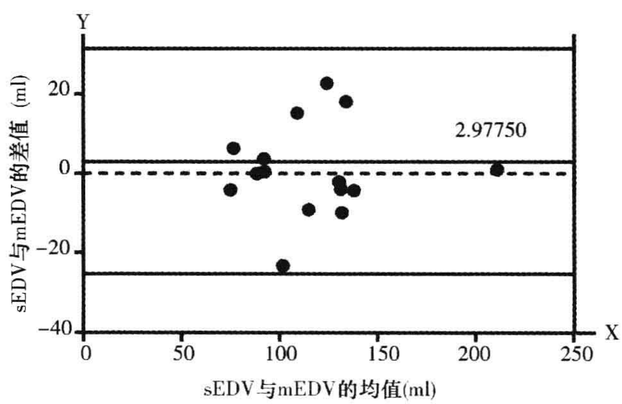
MRI(mEDV)和单次屏气电影法 MRI(sEDV)进行测量,数据见下表。

两种方法测量 16 名检查者 EDV 值(ml)

检查者编号	mEDV	sEDV
1	123.25	86.32
2	126.83	136.65
3	79.7	73.33
4	129.23	133.19
5	110.23	119.34
6	116.53	101.29
7	88.35	88.4
8	90.04	113.25
9	129.32	131.44
10	211.44	210.5
11	142.93	124.82
12	135.41	112.81
13	135.63	139.93
14	92.58	92.13
15	72.85	77.04
16	93.9	90.14

三、该示例的图形表达

在二维坐标系中,横轴 X 表示两种方法测量每个对象的平均值,纵轴 Y 表示两种方法测量每个对象的差值。



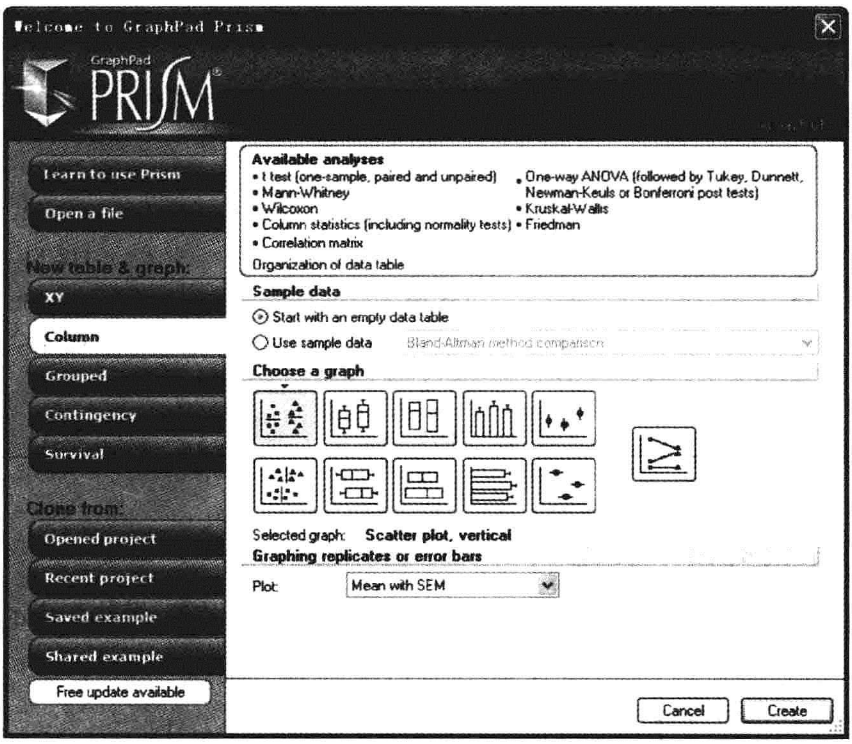
两种方法测量 EDV 的 Bland-Altman 图

1. 偏倚(bias) 当用两种方法对同一批受试对象同时进行测定时,通常不会获得完全一致的结果,而是存在一定差异,如一种方法的测量结果经常大于(或小于)另一种方法的结果,这种差异称为偏倚。偏倚可以用两种方法测量差值的均数来表示,本图中为2.97750。

2. 一致性界限(limits of agreement) 差值均数的变异情况用其标准差 S_d 来描述,如果差值的分布服从正态分布,那么95%的差值应当位于 $\bar{d}-1.96S_d$ 和 $\bar{d}+1.96S_d$ 之间,我们称这个区间为95%的一致性界限,绝大多数差值都位于该区间内。该图一致性界限为-25.3042至31.2592。

四、图形绘制

Step 1:**图形类型选择**



1. 左侧 New table & graph 项选择 Column。
2. 右上 Sample data 项选择 Start with an empty data table,即建立新的数据文件。
3. 右中 Choose a graph 不需要选择,因为图形需重新绘制。

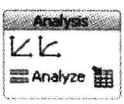
Step 2:**数据录入**

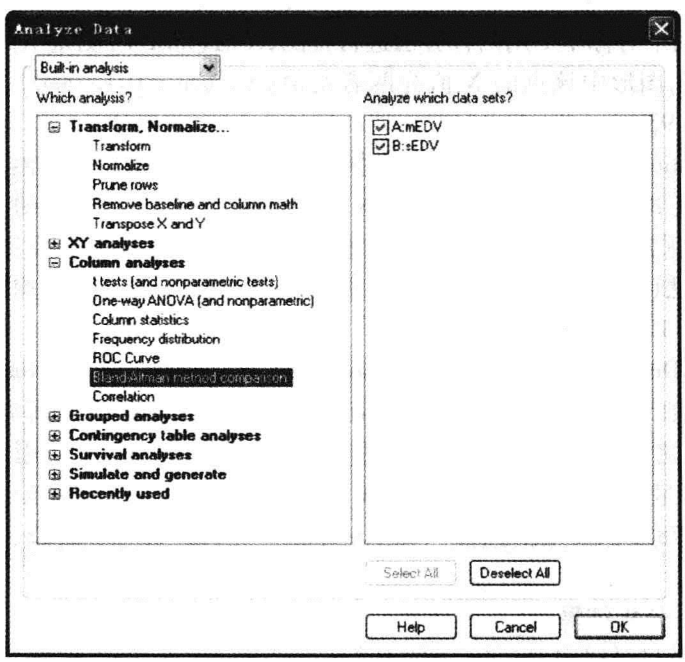
软件自动进入数据录入窗口,即 data 窗口,按下图录入数据。

	A	B
	mEDV	sEDV
	Y	Y
1	123.25	86.32
2	126.83	136.65
3	79.70	73.33
4	129.23	133.19
5	110.23	119.34
6	116.53	101.29
7	88.35	88.40
8	90.04	113.25
9	129.32	131.44
10	211.44	210.50
11	142.93	124.82
12	135.41	112.81
13	135.63	139.93
14	92.58	92.13
15	72.85	77.04
16	93.90	90.14

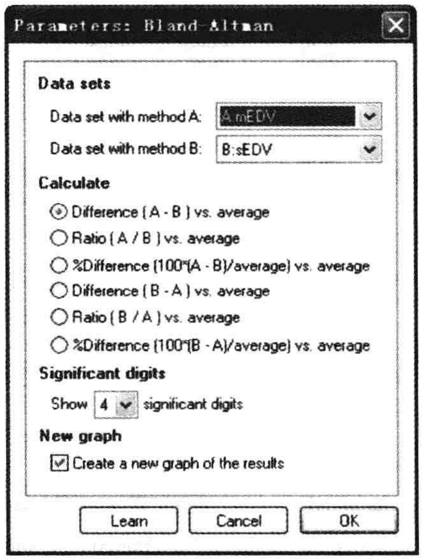
Step 3:**数据分析**

1. 快捷菜单栏中数据分析模块中的分析按钮 Analyze,弹出分析对话框。选择 Column analyses 分析中的 Bland-Altman method comparison,即 Bland-Altman 绘图对话框(Parameters: Bland-Altman)。





2. 这是 Bland-Altman 图形设置中最重要的对话框,其中只需要对 Calculate 选项进行设置即可,本次选择默认选项 Difference(A – B) vs. average。



(1)Difference (A – B) vs. average:其中 average 为 X 轴,即两次检测值的平均值,Y 轴为两次检测的差值 Difference (A – B)。若检测值比较稳定,通常选择该选项,也是运用最多的选项。如示例中对标本 1 用两种方法进行检测,A 法(mEDV)测量值为123.25,B 法(sEDV)测量值为 86.32,图形中该点的 X 值为两者均值(average = 104.7850),Y 值为 A – B = 123.25 – 86.32 = 36.93。

(2)Ratio (A/B) vs. average:其中 average 为 X 轴,即两次检测值的平均值,Y 轴为两次检测值的比值 Ratio(A/B)。如果随着 X 轴数值的增加,两次检测的差值(Difference)也随之增大,则可以采用该选项。如示例中对标本 1 用两种方法进行检测,A 法(mEDV)测量值为 123.25,B 法(sEDV)测量值为 86.32,图形中该点的 X 值为两者均值(average = 104.7850),Y 值为 A/B = 123.25/86.32 = 1.4278。

(3)% Difference(100 * (A – B)/average) vs. average:其中 average 为 X 轴,即两次检测的平均值,Y 轴为两者差值除以两者均值。如果随着 X 轴数值的增加,两次检测的差值(Difference)也随之增大,则可以采用该选项。如示例中对标本 1 用两种方法进行检测,A 法(mEDV)测量值为 123.25,B 法(sEDV)测量值为 86.32,图形中该点的 X 值为两者均值(average = 104.7850),Y 值为(A – B)/average = (123.25 – 86.32)/104.7850 = 0.352436 = 35.2436%。

Step 4:**分析结果**

1. 点击目录树中 Results 下的 Difference vs average,则出现如下结果:

		X	A
		Average	Difference
		X	Y
1		104.7850	36.9300
2		131.7400	-9.8200
3		76.5150	6.3700
4		131.2100	-3.9600
5		114.7850	-9.1100
6		108.9100	15.2400
7		88.3750	-0.0500
8		101.6450	-23.2100
9		130.3800	-2.1200
10		210.9700	0.9400
11		133.8750	18.1100
12		124.1100	22.6000
13		137.7800	-4.3000
14		92.3650	0.4500
15		74.9450	-4.1900
16		92.0200	3.7600

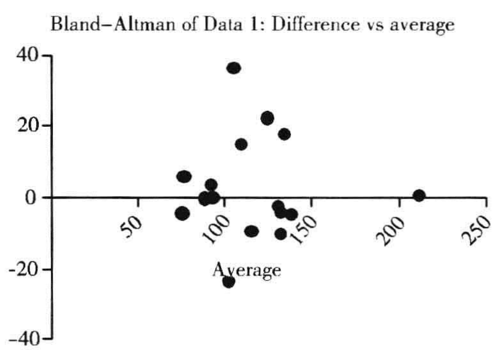
这是 Analyze 对话框中 Difference（A－B）vs. average 所出现的结果，X 轴为 average，Y 轴为 Difference(A－B)，Bland－Altman 图即按照该表进行绘制。

2. 点击目录树中 Results 下的 Bias & agreement，则出现如下结果：

	Bland-Altman Bias & agreement	
1	Bias	2.97750
2	SD of bias	14.4294
3	95% Limits of Agreement	
4	From	-25.3042
5	To	31.2592

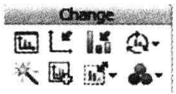
从该表中获得三个最重要的参考值，bias(偏倚)为 2.97750，而 95% Limits of Agreement (95%一致性界限)为(－25.3042, 31.2592)。最后依据这三个值对图形添加三条参考线。

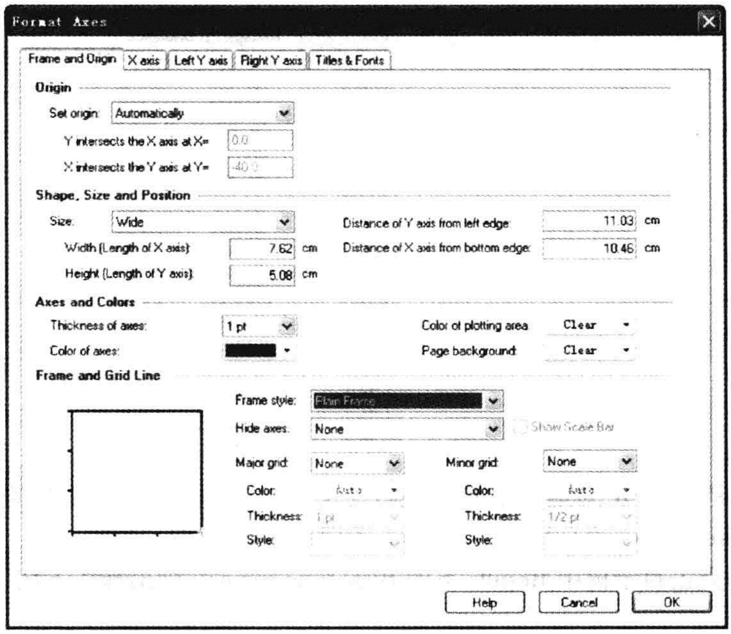
3. 点击目录树中 Graphs 下的 Bland-Altman of Data 1，软件自动生成如下图形：



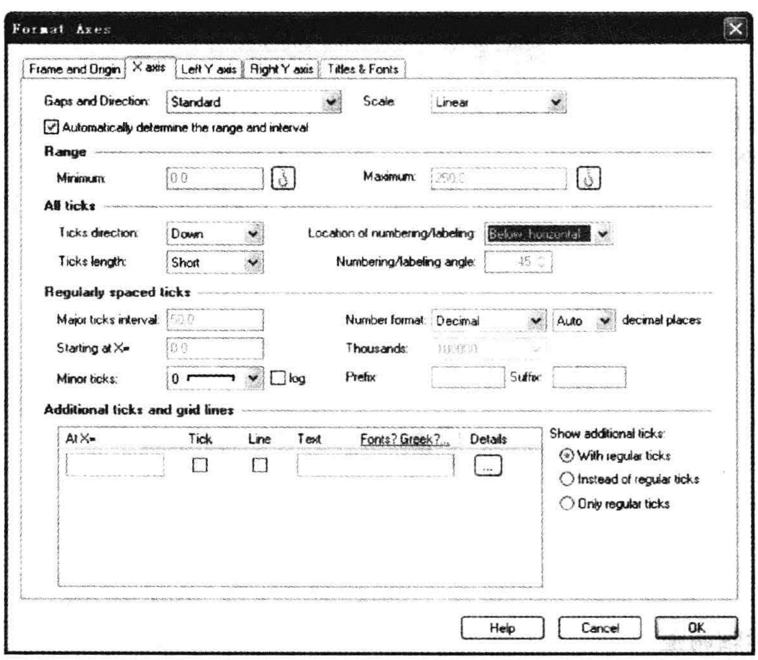
Step 5：**坐标轴调整**

1. 点击软件快捷菜单栏 Change 模块中上左二按钮，弹出坐标轴对话框。选择 Frame and Origin，对坐标系进行调整。在 Frame and Grid Line 中对 Frame style 项进行选择，选中 Plain Frame，具体式样在左下角图示。



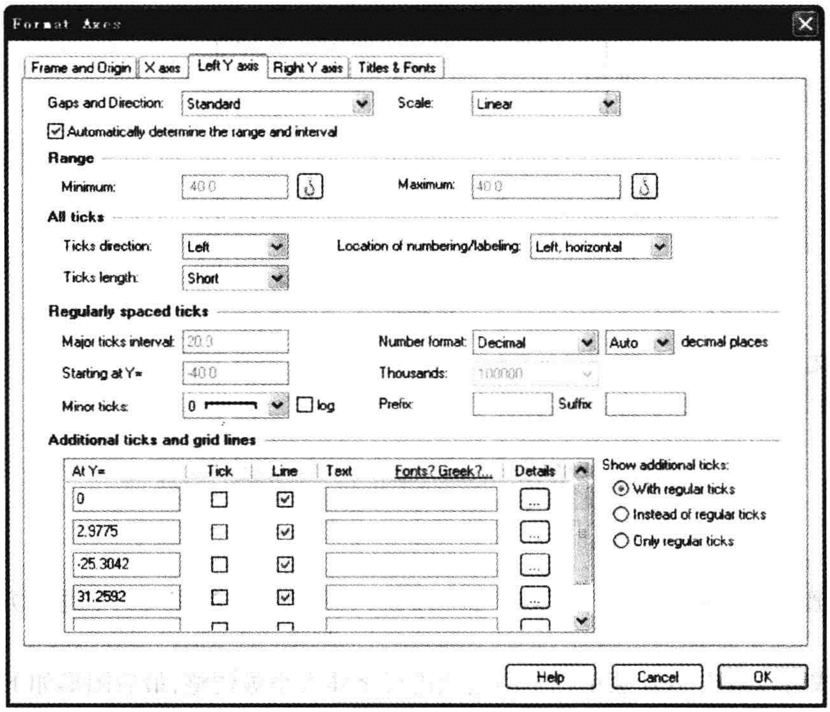


2. 选择 X axis，对 X 轴进行调整。在 All ticks 中对 Location of numbering/labeling 项进行选择，选中 Below，horizontal，即 X 轴刻度文字出现在轴下方，横向排列。



3. 选择 Left Y axis，添加四条参考线。由于 Plain Frame 坐标系中 X 轴处于坐标系的下方，故需要添加 Y = 0 的参考线，在 Additional ticks and grid lines 下填写 At Y = 0，并选中 Line 框，如下图所示。添加偏倚参考线 Y = 2.9775。添加 95% 一致性界限下限 Y = −25.3042。添

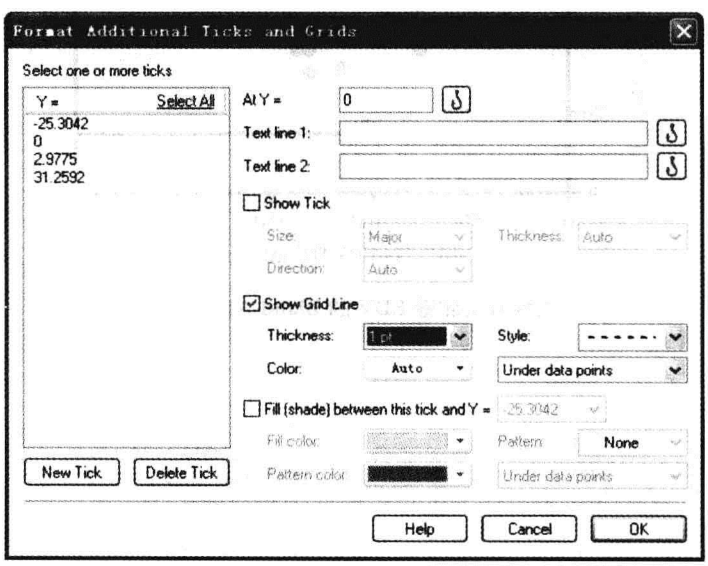
加 95%一致性界限上限 Y = 31.2592。点击右侧的 Details 按钮,弹出 Format Additional Ticks and Grids 对话框,对参考线的式样进行调整。



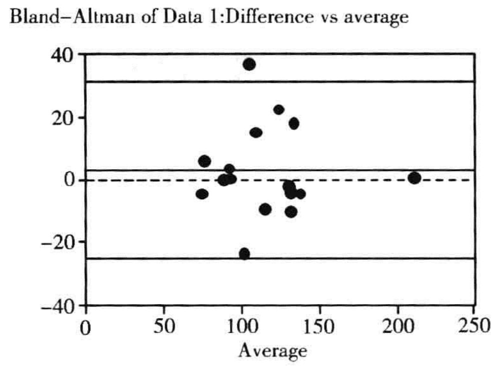
4. 在左边对话框中,选中 0,即对 Y = 0 参考线进行调整

在 Show Grid Line 下调整 Thickness: 1 pt, Style: - - - -

依次对其他参考线进行调整

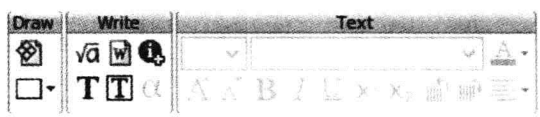


5. 最后形成如下图形：



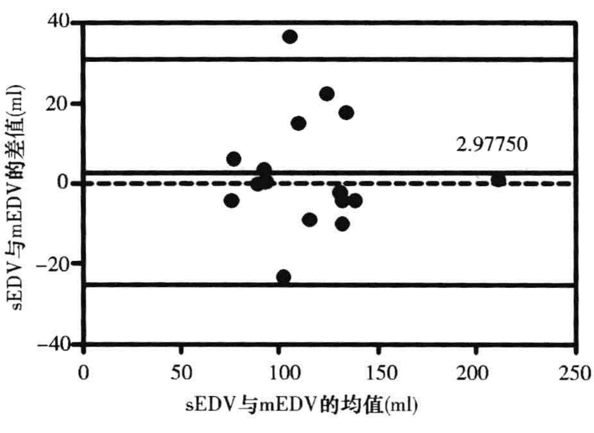
Step 6：**文字修饰**

将图形中不需要的文字点击删除即可。



点击快捷菜单栏 Write 中的 T(文字输入按钮)，然后在想要输入文字的地方输入文字即可。

输入完成可以点击 Text 菜单栏中的选项进行字体大小等调整，最后图形如下：



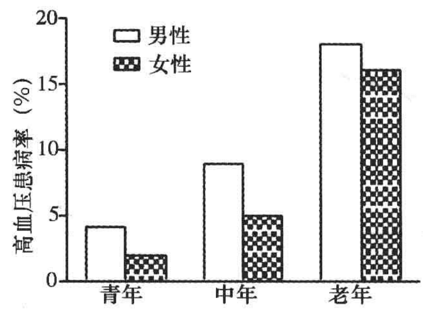
两种方法测量 EDV 的 Bland-Altman 图

下篇

统计绘图篇(GraphPad Prism 5)

>> 第十四章　统计图基本知识

一、统计图基本要素



1990 年我国城市部分年龄段不同性别的高血压患病率

1. 标题　标题一般位于图的正下方,简单扼要地说明图形要表达的主要内容。

2. X 轴　X 轴有两种,一种为分类轴,即 X 轴表示不同的类别,另一种为数据轴,即 X 轴表示一个数值概念,每个刻度具有特定的意义,刻度可以用一般数值或对数值表示。

3. Y 轴　Y 轴一般为数值轴,表示测量关注的结果。有时可能出现双 Y 轴,即左右 Y 轴。如果存在双 Y 轴,则要注意各 Y 轴所对应的具体图形。

4. 图形区　即 X 轴和 Y 轴所组成的二维平面内,用点、线和直条表达数据,统计图和统计表是相对应的,图形区内的点、线和直条的位置和形状均有相应的数据意义。

二、几个必须弄懂的问题

(一)统计图的关键在于“合适”和“规范”

统计图是用图形将数据形象直观地表示出来,各个绘图软件都宣称绘图功能多样,但在医学科研中,统计图也就是简单的散点图、线图、条图和面积图这样几类,关键在于表达“合适”,如成组 t 检验和配对 t 检验两者的图形表达就有差别,统计图的绘制不是进行艺术创作,而是依据统计方法,选用相应的图形,而且图形必须规范,四个基本要素缺一不可。

(二)不是所有的数据均可以用图形表达

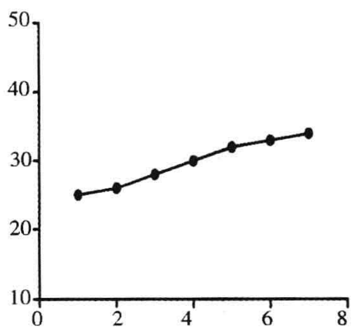
虽然统计图是为了表达统计数据,但是很多统计方法却不能用统计图来表达,如多重线性回归,所以我们能够用统计图来表达的统计数据非常有限,本书以统计方法为序,向大家介绍常用的统计图形的绘制。

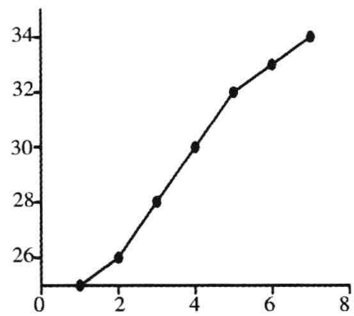
（三）不需要将所有的统计结果均用图形表达

大家常常有这样的冲动，即如果该统计分析结果能够用图形表达，就会绘制出相应的图形。记住，统计图只是为了说明非常重要的统计结果，尤其是论文中的首要结局（primary outcome）。但有些统计数据，如临床试验中的基线资料，由于结果并不重要，一般只需列出统计表格，不需要绘制统计图。

（四）图形有时可以误导读者

虽然“一图值千字”，而且图形有直观形象的优点，但是同一组数据通过图形的细微改变，就会给读者带来不同的印象，如下面两张图：

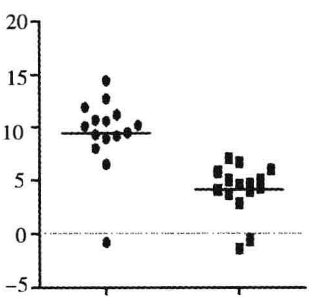


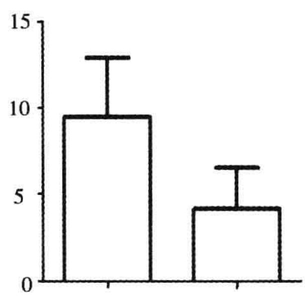


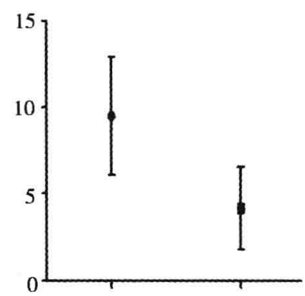
两图来自同一组数据，但仅仅 Y 轴的刻度范围不同，从而给读者造成迥然不同的印象。

三、常见统计方法的图形表达

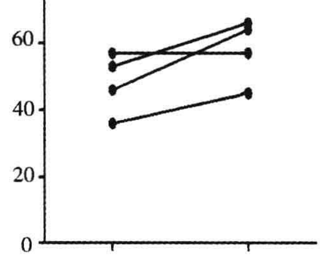
1. 成组 t 检验



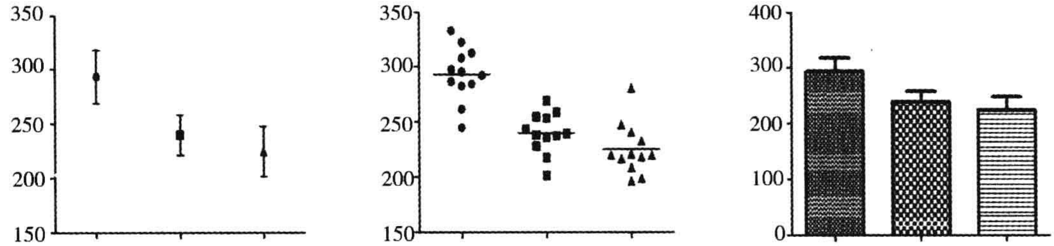




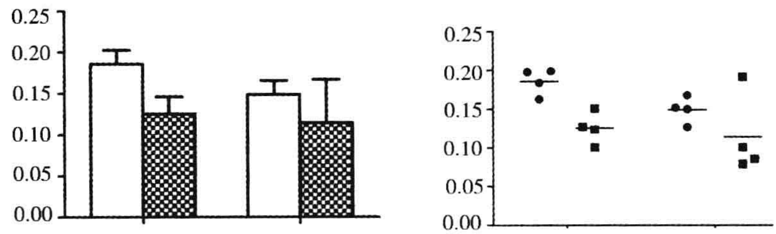
2. 配对 t 检验



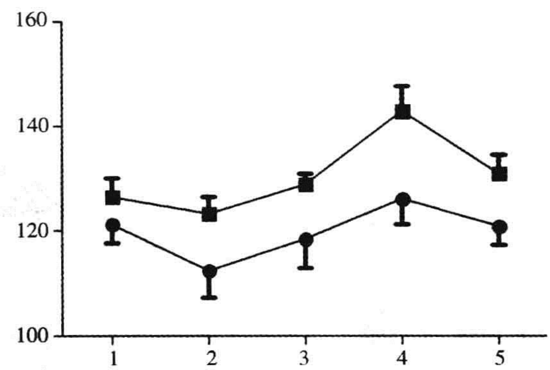
3. 完全随机设计资料的方差分析



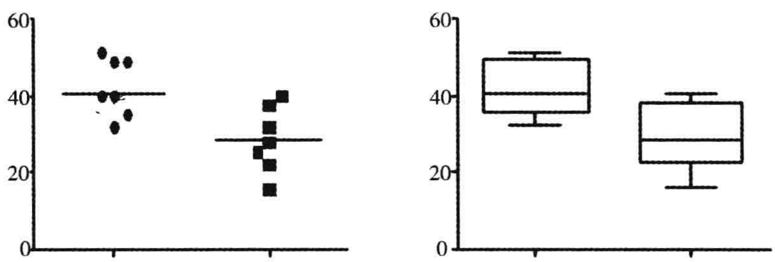
4. 析因设计资料的方差分析



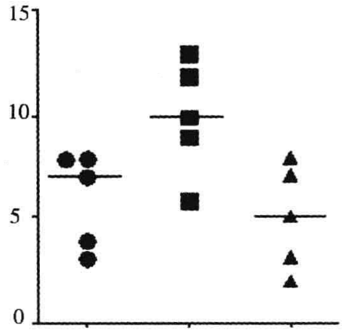
5. 重复测量资料的方差分析

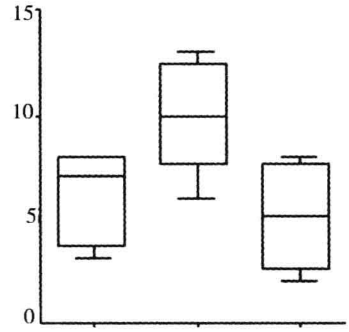


6. 两组独立样本秩和检验

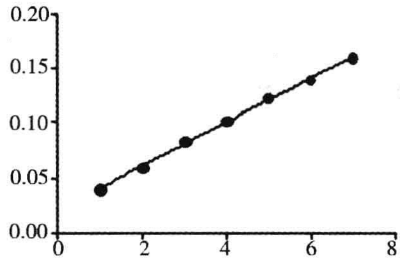


7. 多组独立样本秩和检验

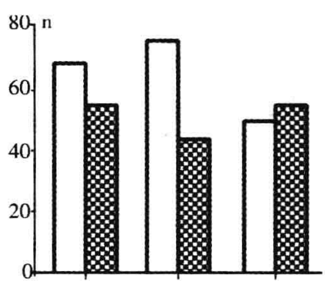


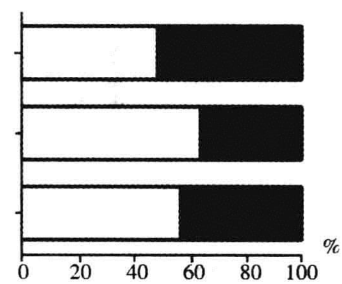


8. 简单直线回归和线性相关

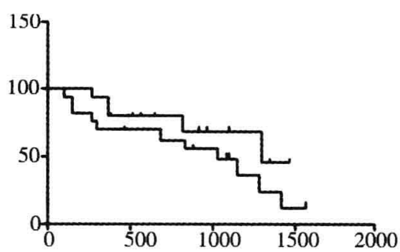


9. 列联表分析





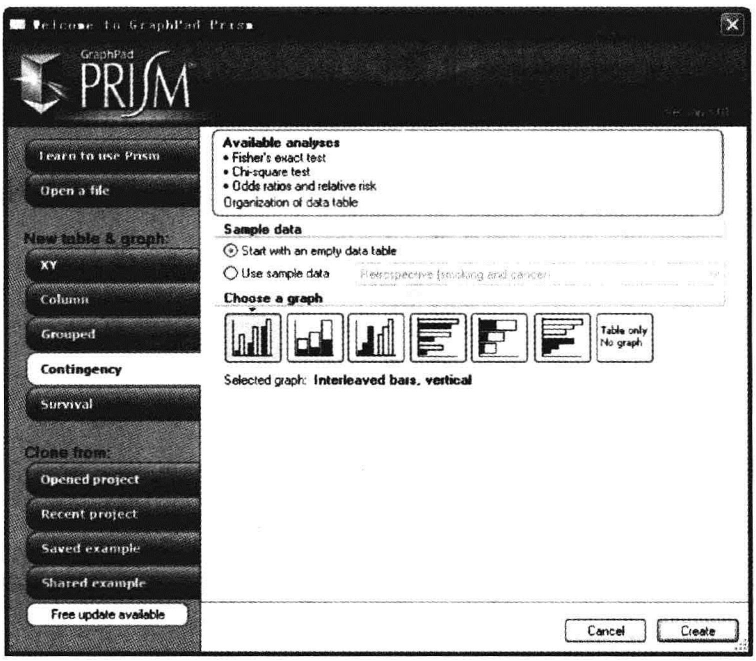
10. 生存分析



第十五章　GraphPad Prism 5 绘图界面介绍

一、欢迎界面

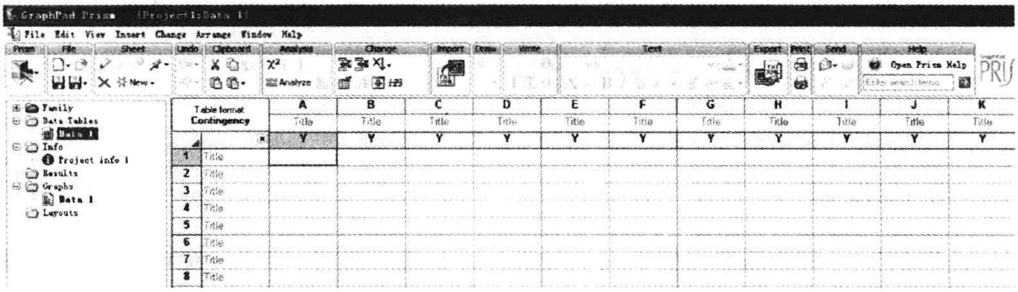
GraphPad Prism 的欢迎界面要求大家对统计方法进行选择,特定的统计方法对应相应的数据录入格式和图形类型,如 choose a graph 下面缩略图表示对应的图形类型,有时需要设定数据输入格式。



统计类别	统计分析方法
XY	线性回归、非线性回归、相关分析
Column	单样本 t 检验、配对 t 检验、成组 t 检验、单样本秩和检验、两组独立样本秩和检验、多组独立样本秩和检验、单因素方差分析
Grouped	两因素方差分析、重复测量的两样本方差分析
Contingency	Fisher 精确检验、卡方检验
Survival	单因素生存曲线比较

二、软件主界面

在欢迎界面对统计类别、图形类型和数据输入格式进行设定后,我们进入软件主界面。



三、目录树

1. Data Tables　Data 1(数据集):用鼠标点击 data 时,则进入数据录入窗口,我们在该窗口录入数据,根据你在欢迎界面设定不同,数据集的表格会有所不同。

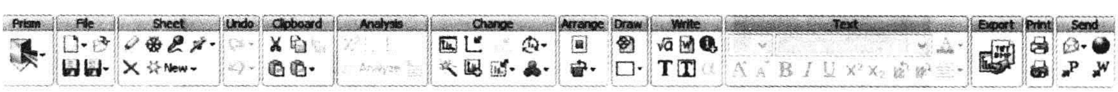
2. Graph　Data 1(图形):当你完成数据录入后,用鼠标点击该图标,则自动生成图形,我们只需要对图形稍加修饰即可。

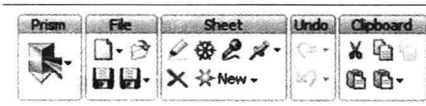
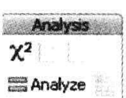
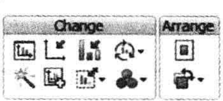
3. Layouts(布局)　在实际论文撰写中,可能会产生多个图表,我们需要对此进行排版。



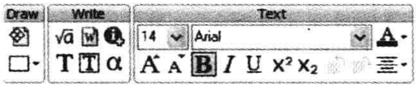
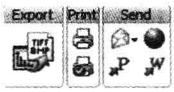
四、快捷菜单栏

当鼠标点击目录树中 Graph: Data 1 时,软件上方出现如下快捷菜单栏:

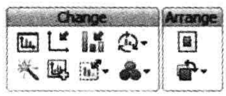


菜单栏	功能
	文档操作模块,如打开、新建、保存、前进、后退、复制和粘贴等,该模块与 office 软件操作一致。
	分析模块,不过由于本书不涉及用 GraphPad Prism 进行统计分析,所以不作讲解。
	非常重要的图形模块,将对图形类型、坐标轴和图形本身进行设定,以下将详细解释。

续表

菜单栏	功能
	文字修饰模块,可以绘制线条、添加文字,并对文字进行调整。
	图形输出模块,包括图形导出、打印和一键输出至word 和 PPT 中。

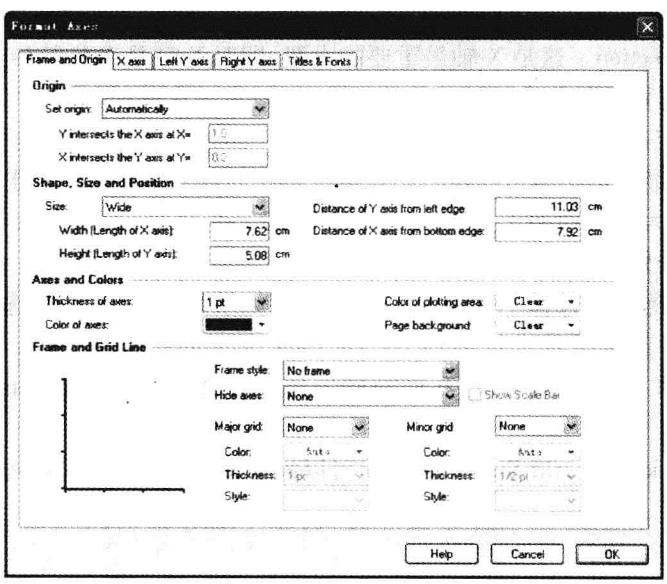
五、图形模块详解



缩略图	功能
	相当于欢迎界面中的统计方法和图形类型选择项,可以对统计方法重新选择,但一般只对图形类型进行选择。
	对坐标轴进行具体设定,包括坐标系、X 轴和 Y 轴。
	对图形进行具体调整。
	图表的配色方案,默认为黑白图表,但若发表在网上或制作幻灯片,则需要对此进行选择。

六、坐标轴调整

(一) Frame and Origin(坐标框和原点)



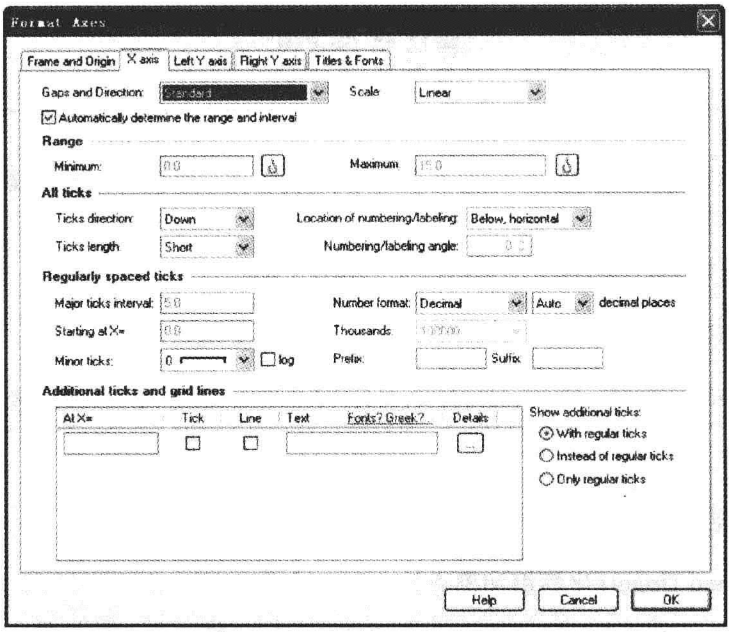
1. Origin 原点设定，即 X 轴与 Y 轴交叉点，默认为软件自动设定，但不一定交汇于(0,0)，有 lower left(左下)、upper left(左上)、lower right(右下)、upper right(右上)和 custom(自定义)几个选项。

2. Shape, Size and Position 绘图区的形状、大小和位置进行设定，此处一般不需要设定。

3. Axes and Colors　Thickness of axes 和 Color of axes 是对坐标轴的粗细和颜色进行设定，Color of plotting areas 对绘图区的颜色进行设定，Page background 对背景颜色进行设定。

4. Frame and Grid line　是对图形框和网格线进行设定，由于此处左侧有图示，大家设定时尝试一下即可。

(二)X axis(X 轴)

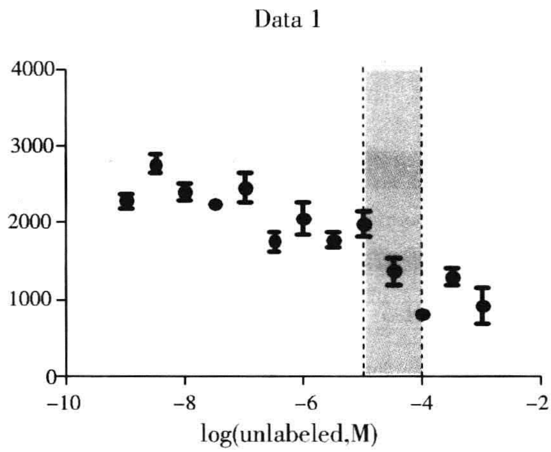


1. Gaps and Direction　这是 X 轴最重要的设置，即对 X 轴是否分段和方向进行设定(值得注意的是，只有在 XY 图形中才可以对此进行设定，因为只有此时 X 轴才为数据轴，而非分类轴)。如果需要设定，则将 Automatically determine the range and interval 前面的勾去除，同时在下面 Range 项中对最大最小值进行设定。此处有四个选项：Standard(标准)、Reverse(逆向，即 X 轴从左到右为从大到小排列)、Two segments(两段)和 Three segments(三段)，如果选择分段，还需要对各段分别进行设定。

2. Scale　对坐标轴的选择一般为 linear(线性)，如果为浓度，则一般选择 log(对数)刻度。

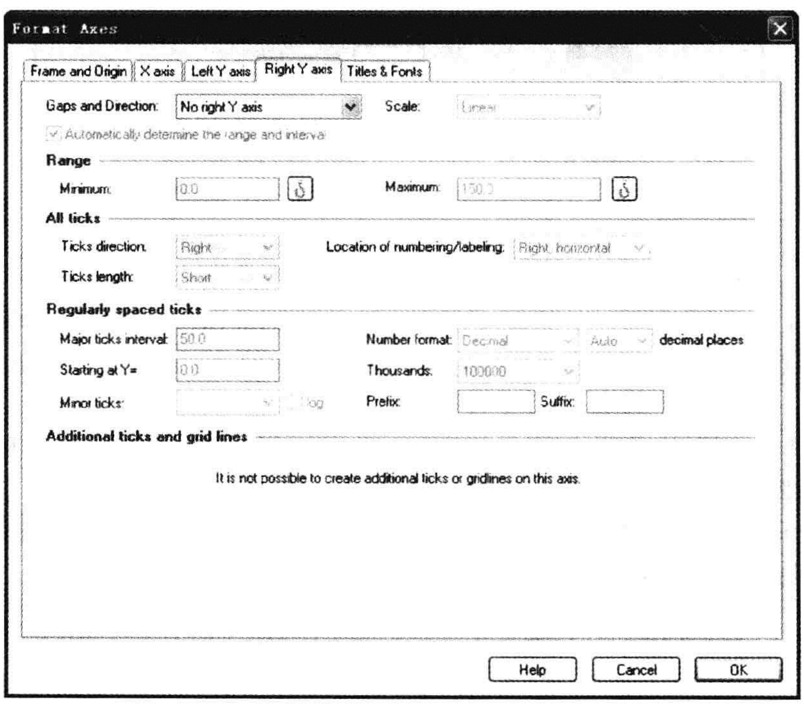
3. All ticks　对刻度图示进行调整，如 Location of numbering/labeling 选项默认为 below, angle(刻度文字在 X 轴下方，倾斜)，我们需要将其调整为 Below, horizontal(刻度文字在 x 轴下方，水平)

4. Additional ticks and grid lines　该选项可以添加辅助线，还可以对两条辅助线内的区域进行填色，见下图：



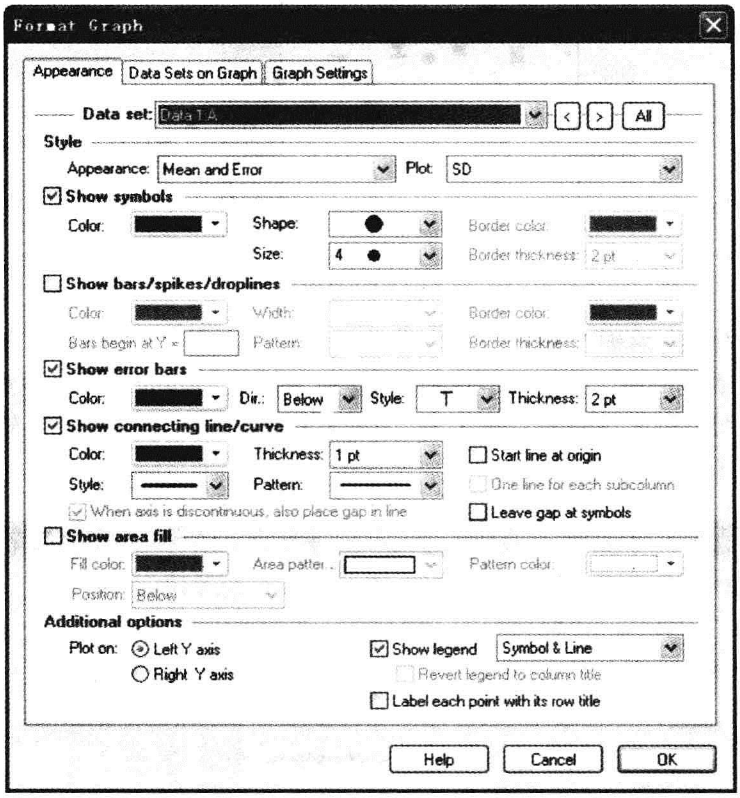
(三)Right Y axis(右侧 Y 轴)

软件默认无右侧的第二条 Y 轴,如果需要绘制,则在此处设定。需要注意的是,添加第二条 Y 轴后,所有图形仍旧均跟随左侧 Y 轴,如果某些数据需要按照右侧 Y 轴,需要在 Format Graph(图形设定)对话框中实现。



七、图形设定界面





1. Data set 数据集：这是一个非常重要的概念，即图形的操作是针对单一数据集（也就是变量）而言的，因此可以对图形分别进行调整。如果选择 Change all data sets，则对所有数据集进行调整。

2. Style：样式，不同的图形，此处出现不同选项，当选用不同的 Appearance 和 Plot 后，在下方的各选项对图形进一步调整。

3. Additional options（其他选项）中的 Plot on 可以选择按数据集是否依据左侧 Y 轴还是右侧 Y 轴进行显示，如果出现第二 Y 轴，此处应进行设定。

>> 第十六章　各种统计方法所对应的统计图绘制

第一节　配对 t 检验的图形绘制

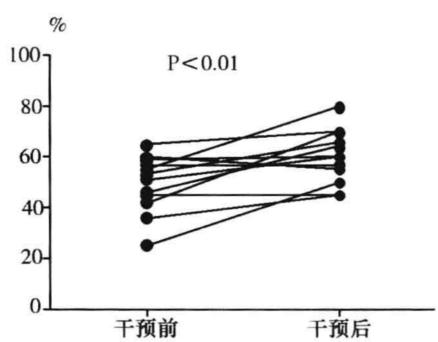
一、示例

某地区随机抽取 12 名贫血儿童家庭,实行健康教育干预 3 个月,干预前后儿童的血红蛋白(%)测量结果如下表所示,试问干预前后该地区贫血儿童血红蛋白(%)平均水平有无变化?(《卫生统计学》第六版例 7-2,P104)。

干预前血红蛋白	36	46	53	57	65	60	42	45	25	55	51	59
干预后血红蛋白	45	64	66	57	70	55	70	45	50	80	60	60

二、该示例的图形表达

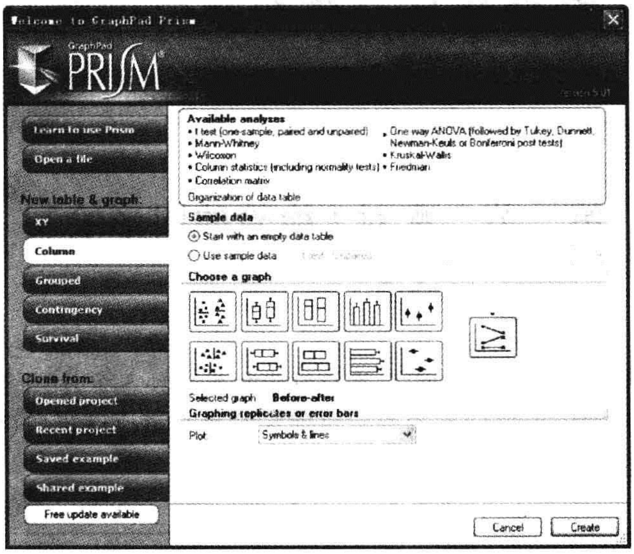
配对 t 检验图形表达的形式应与成组 t 检验区别开来,配对 t 检验应当采用单一研究对象前后值连线的方式表达,但如果数据太多,则不适合用图形表达(如下图)。



干预前后贫血儿童蛋白质改变

三、图形绘制

Step 1:图形类型选择



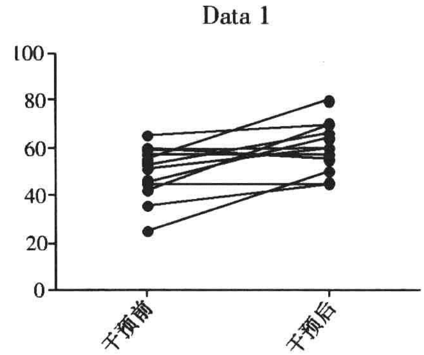
1. 左侧 New table & graph 项选择 Column,该类型适合常见的 t 检验、方差分析和秩和检验。
2. 右上 Sample data 项选择 start with an empty data table,即建立新的数据文件。
3. 右中 Choose a graph 中选择最左侧的 Before-after 图形类型。
4. 右下 Graphing replicates or error bars 中选择 symbols & lines,即绘制各点并连线表示。

Step 2:数据录入

1. 软件自动进入数据录入窗口,即 data 窗口,按下图录入数据。

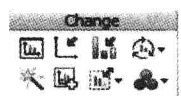
Table format: Column		A 干预前	B 干预后
		Y	Y
1	Title	36	45
2	Title	46	64
3	Title	53	66
4	Title	57	57
5	Title	65	70
6	Title	60	55
7	Title	42	70
8	Title	45	45
9	Title	25	50
10	Title	55	80
11	Title	51	60
12	Title	59	60

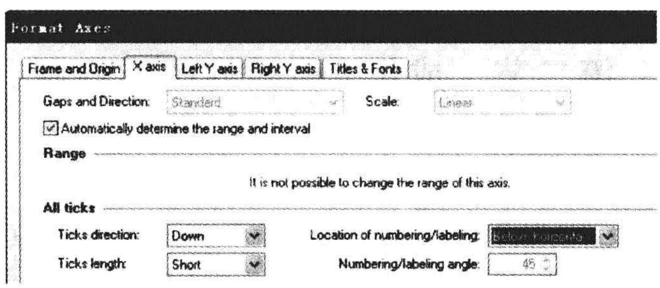
2. 点击左侧目录树中的 graphs 图标,可见自动生成如下图形。



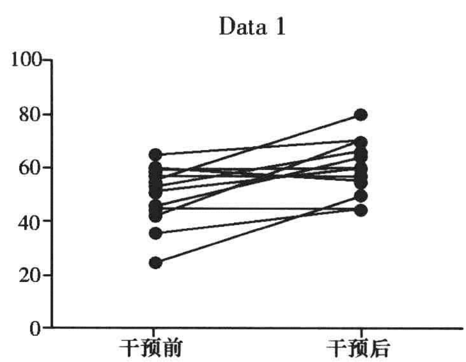
Step 3:**坐标轴和图形调整**

点击软件快捷菜单栏 change 模块中上左二按钮,弹出坐标轴对话框。





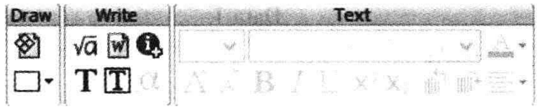
选择 X axis,对 X 轴进行调整。在 All ticks 中对 Location of numbering/labeling 项进行选择,选中 Below, horizontal,即 X 轴刻度文字出现在轴下方,横向排列,图形如下:

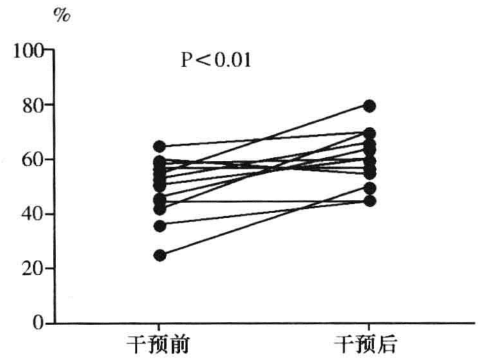


Step4:**文字修饰**

将图形中不需要的文字点击删除即可。点击快捷菜单栏 Write 中的 T(文字输入按钮),然后在想要输入文字的地方输入文字即可,输入完成可以点击 Text 菜单栏中的选项进行字体

大小等调整，最后图形如下：





干预前后贫血儿童蛋白质改变

第二节　成组 t 检验的图形绘制

一、示例

某妇产医院的研究者欲探讨孕妇在孕期补充钙剂对血清骨钙素(ng/ml)的影响，选取孕妇的年龄，基础骨钙素值接近，孕周在 26～28 周的 30 名孕妇，随机分成两组，每组 15 人。试验组孕妇补充选定的某种钙制剂，对照组孕妇采用传统膳食，产后 40～50 天内测定两组孕妇血清骨钙素的改变值，结果如下：(《卫生统计学》第六版例 7-4，P106)

试验组	10.2	8.9	10.1	9.2	-0.8	10.6	6.5	11.2	9.3	8.0	10.7	9.5	12.7	14.4	11.9
健康人	5.0	6.7	-1.4	4.0	7.1	-0.6	2.8	4.3	3.7	5.8	4.6	6.0	4.1	5.1	4.7

二、该示例的图形表达

1. 散点图(图 1)　清晰地描述了各点数据的情况，散点图容易发现离群值，散点中的横线为均数(mean)。当数据较少时用散点图表达比较合适。

2. 误差图(图 2)　箱顶为均数(mean)，触须线表示标准差(SD)，也可以表示标准误(SEM)，但表示标准误可能误导读者。当数据较多时用误差图表达比较合适。

3. 误差图(图 3)　点表示均数，直线表示两侧标准差(SD)。

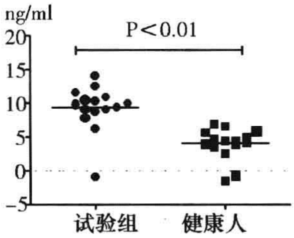


图1　两组孕妇骨钙素改变值

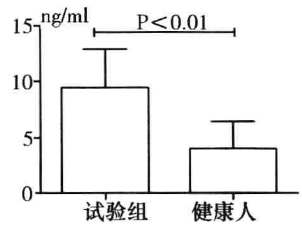


图2　两组孕妇骨钙素改变值

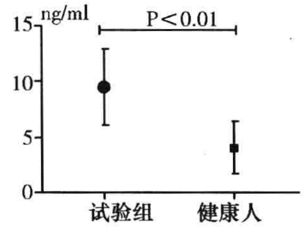
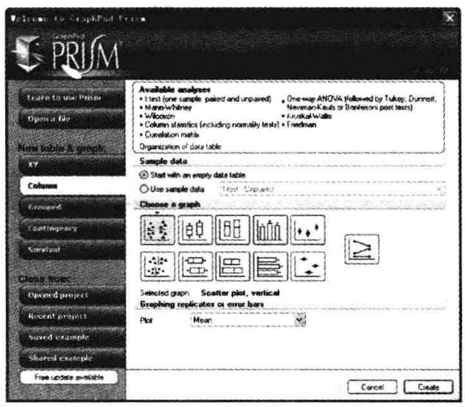


图3　两组孕妇骨钙素改变值

三、图形绘制

Step 1:图形类型选择

1. 左侧 New table & graph 项选择 Column,该类型适合常见的 t 检验、方差分析和秩和检验。
2. 右上 Sample data 项选择 Start with an empty data table,即建立新的数据文件。
3. 右中 Choose a graph 中选择左上的 Scatter plot, vertical 图形类型,即纵向散点图。
4. 右下 Graphing replicates or error bars 中选择 Mean,即用横线表示均数。

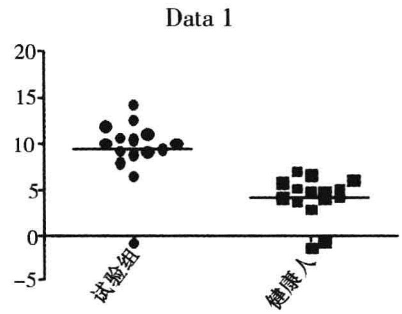


Step 2:数据录入

软件自动进入数据录入窗口,即 data 窗口,按下图录入数据。

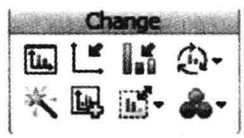
	A	B
	试验组	健康人
	Y	Y
1	10.2	5.0
2	8.9	6.7
3	10.1	-1.4
4	9.2	4.0
5	-0.8	7.1
6	10.6	-0.6
7	6.5	2.8
8	11.2	4.3
9	9.3	3.7
10	8.0	5.8
11	10.7	4.6
12	9.5	6.0
13	12.7	4.1
14	14.4	5.1
15	11.9	4.7

点击左侧目录树中的 graphs 图标,可见自动生成如下图形。

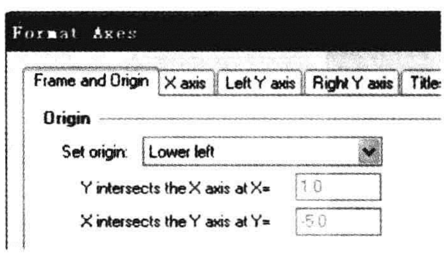


Step 3:**坐标轴和图形调整**

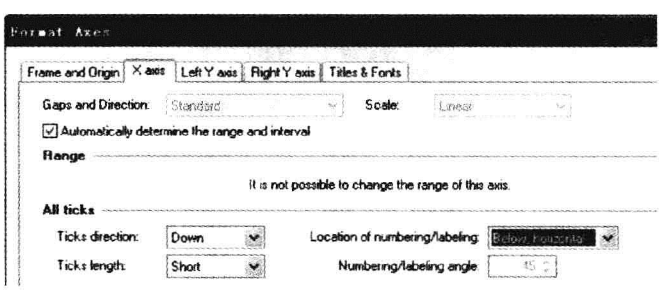
点击软件快捷菜单栏 Change 模块中的上左二按钮,弹出坐标轴对话框。



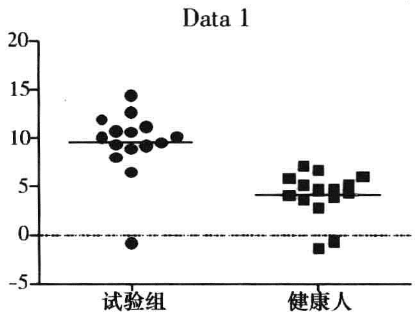
选择 Frame and Origin,将对整体坐标轴进行调整。在 Origin 中对 Set origin 项进行选择,选中 Lower left,即坐标原点的位置出现在左侧 Y 轴最下方。



选择 X axis,对 X 轴进行调整。

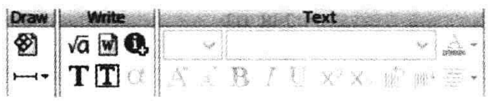


在 All ticks 中对 Location of numbering/labeling 项进行选择,选中 Below, horizontal,即 X 轴刻度文字出现在轴下方,横向排列,图形如下:

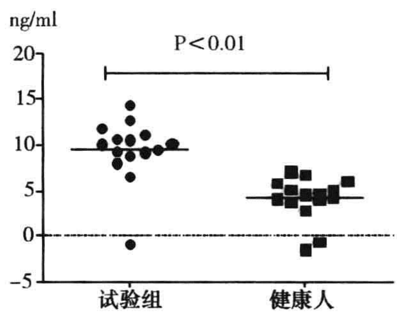


Step4:**文字修饰**

将图形中不需要的文字选择删除即可。点击快捷菜单栏 Draw 中的直线输入按钮,在合适位置绘制直线。



点击快捷菜单栏 Write 中的 T(文字输入按钮),然后在想要输入文字的地方输入文字即可,输入完成可以点击 Text 菜单栏中的选项进行字体大小等调整,最后图形如下:



两组孕妇骨钙素改变值

第三节　完全随机设计资料方差分析的图形绘制

一、示例

为研究钙离子对体重的影响作用,某研究者将 36 只肥胖模型大白鼠随机分为 3 组,每组 12 只,分别给予高脂正常剂量钙(0.5%),高脂中剂量钙(1.0%)和高脂高剂量钙(1.5%)三种不同的饲料,喂养 9 周,测其喂养前后体重的差值。问三组不同喂养方式下的大白鼠体重改变是否不同?(《卫生统计学》第六版例 8-1,P123)

三种不同喂养方式下大白鼠体重喂养前后差值

正常钙(0.5%)	中剂量钙(1.0%)	高剂量钙(1.5%)
332.96	253.21	232.55
297.64	235.87	217.71
312.57	269.3	216.15
295.47	258.9	220.72
284.25	254.39	219.46
307.97	200.87	247.47
292.12	227.79	280.75
244.61	237.05	196.01
261.46	216.85	208.24
286.46	238.03	198.41
322.49	238.19	240.35
282.42	243.49	219.56

二、该示例的图形表达

1. 误差图(图 1)　点表示均数,直线表示两侧的标准差(SD)。当数据较多时适合用误差图表示。

2. 散点图(图 2)　清晰地描述了各点数据的情况,容易发现离群值,散点中的横线为均数(mean)。当数据较少时适合用散点图表示。

3. 误差图(图 3)　箱顶为均数(mean),触须线表示标准差(SD)。

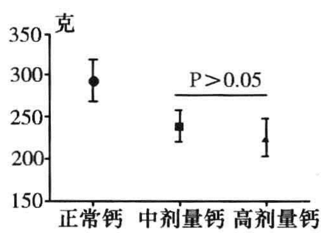


图 1　三组大白鼠体重改变

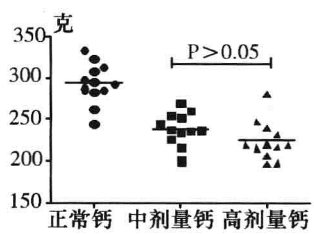


图 2　三组大白鼠体重改变

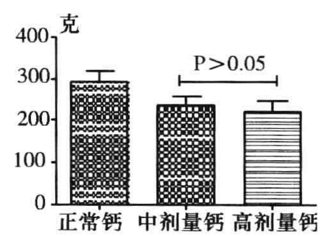


图 3　三组大白鼠体重改变

三、图形绘制

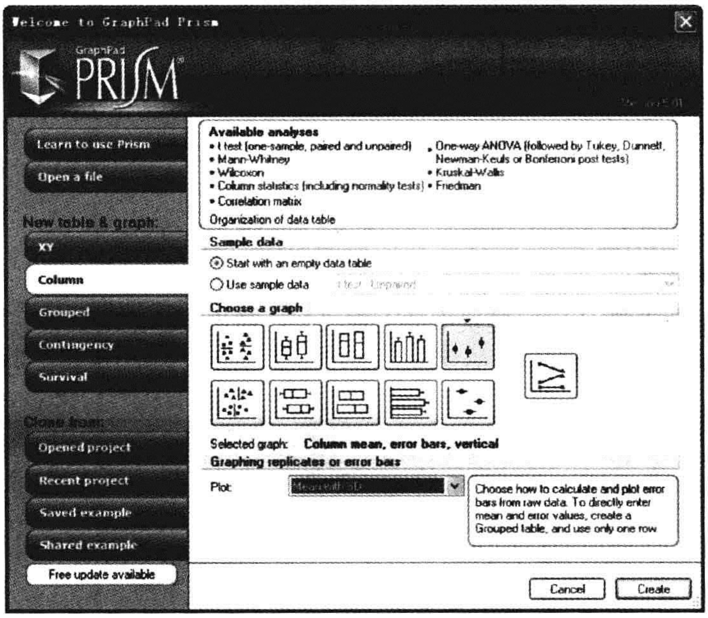
Step 1:图形类型选择

1. 左侧 New table & graph 项选择 Column,该类型适合常见的 *t* 检验、单因素方差分析和秩和检验。

2. 右上 Sample data 项选择 start with an empty data table,即建立新的数据文件。

3. 右中 Choose a graph 中选择左上的 Column mean, error bars, vertical 图形类型,即纵向误差图。

4. 右下 Graphing replicates or error bars 中选择 Mean with SD,即表示均数和标准差。

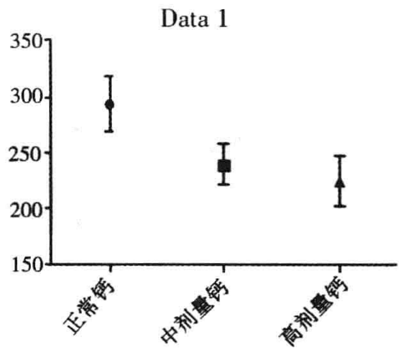


Step 2:**数据录入**

软件自动进入数据录入窗口,即 data 窗口,按下图录入数据。

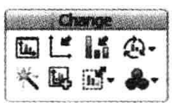
	A	B	C
	正常钙	中剂量钙	高剂量钙
	Y	Y	Y
1	332.96	253.21	232.55
2	297.64	235.87	217.71
3	312.57	269.30	216.15
4	295.47	258.90	220.72
5	284.25	254.39	219.46
6	307.97	200.87	247.47
7	292.12	227.79	280.75
8	244.61	237.05	196.01
9	261.46	216.85	208.24
10	286.46	238.03	198.41
11	322.49	238.19	240.35
12	282.42	243.49	219.56

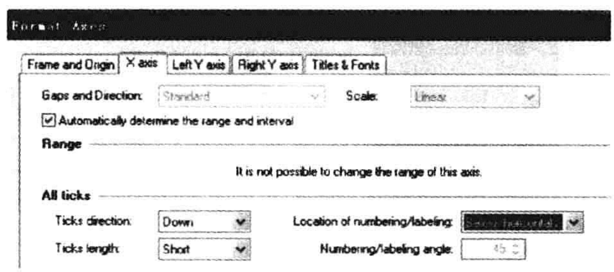
点击左侧目录树中的 graphs 图标,可见自动生成如下图形。



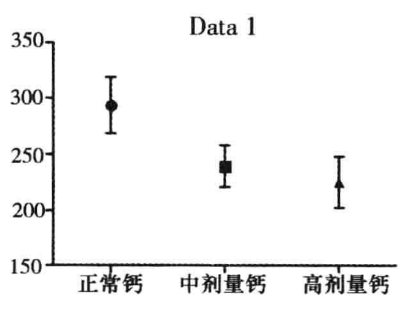
Step 3:**坐标轴调整**

点击软件快捷菜单栏 change 模块中的上左二，弹出坐标轴对话框。
选择 X axis，对 X 轴进行调整。



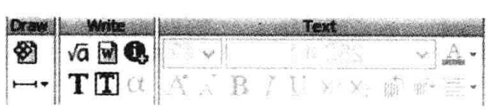


在 All ticks 中对 Location of numbering/labeling 项进行选择，选中 Below，horizontal，即 X 轴刻度文字出现在轴下方，横向排列，图形如下：



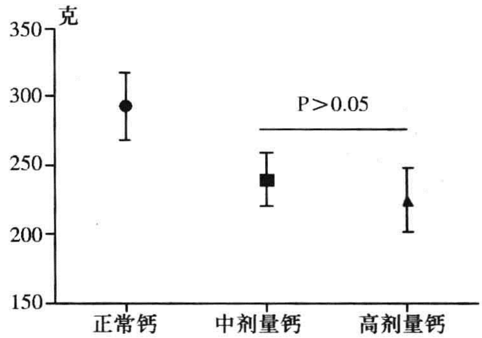
Step4:**文字修饰**

将图形中不需要的文字点击删除即可。点击快捷菜单栏 Draw 中的直线输入按钮，在合适位置绘制直线。



点击快捷菜单栏 Write 中的 T(文字输入按钮)，然后在想要输入文字的地方输入文字即

可,输入完成后可以点击 Text 菜单栏中的选项进行字体大小等调整,最后图形如下:



三组大白鼠体重改变

第四节　析因设计资料方差分析的图形绘制

一、示例

研究者欲研究煤焦油(因素 A)以及作用时间(因素 B)对细胞毒性的作用,煤焦油含量分别为 3 μg/ml(a_1)和 75 μg/ml(a_2)两个水平,作用时间分别为 6 小时(b_1)和 8 小时(b_2)。将统一制备的 16 盒已培养好的细胞随机分为四组,分别接受 A、B 不同组合情况下的四种处理(a_1b_1、a_1b_2、a_2b_1和 a_2b_2),测得处理液吸光度的值(%),结果如下:(《卫生统计学》第六版例 8-3,P129)

四种不同处理情况下吸光度的值(%)

煤焦油(3 μg/ml)a_1		煤焦油(75 μg/ml)a_1	
时间(6 小时) b_1	时间(8 小时) b_2	时间(6 小时) b_1	时间(8 小时) b_2
0.163	0.127	0.124	0.101
0.199	0.168	0.151	0.192
0.184	0.152	0.127	0.079
0.198	0.150	0.101	0.086

二、该示例的图形表达

1. 误差图(图 1)　箱顶为均数(mean),触须线表示标准差(SD)。当数据较多时适合用误差图表示。

2. 散点图(图 2)　清晰地描述了各点数据的情况,容易发现离群值,散点中的横线为均

数(mean)。当数据较少时适合用散点图表示。

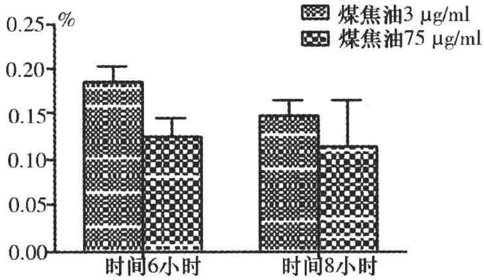


图1 四种不同处理情况下吸光度的值

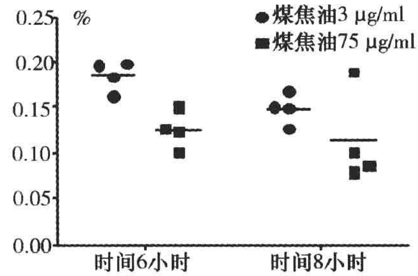
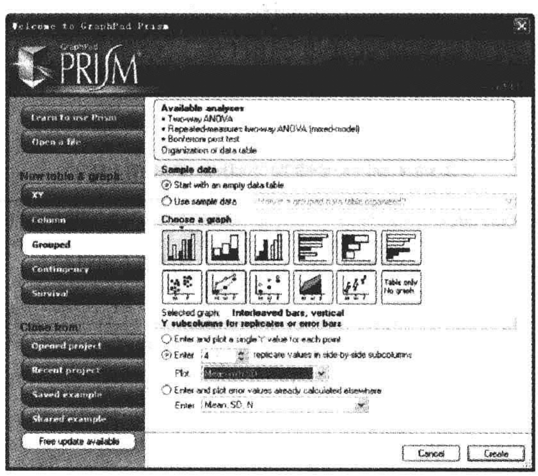


图2 四种不同处理情况下吸光度的值

三、图形绘制

Step 1:图形类型选择

1. 左侧 New table & graph 项选择 Grouped,该类型适合两因素方差分析。
2. 右上 Sample data 项选择 Start with an empty data table,即建立新的数据文件。
3. 右中 Choose a graph 中选择左上的 Interleaved bars, vertical 图形类型,即纵向误差图。
4. 右下 Y subcolumns for replicates or error bars 中选择 Enter 4 replicates values in side-by-side subcolumns,即每个单元格中重复 4 次,如果各单元格重复次数不一致,则填写最大值;Plot 选择 Mean with SD,即表示均数和标准差。

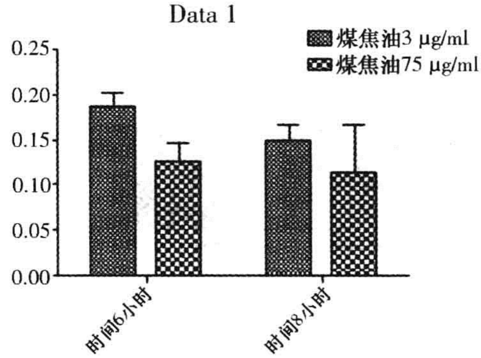


Step 2:数据录入

软件自动进入数据录入窗口,即 data 窗口,按下图录入数据。

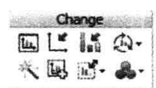
Table format: Grouped	A				B			
	煤焦油3μg/ml				煤焦油75μg/ml			
	A:Y1	A:Y2	A:Y3	A:Y4	B:Y1	B:Y2	B:Y3	B:Y4
1 时间6小时	0.163	0.199	0.184	0.198	0.124	0.151	0.127	0.101
2 时间8小时	0.127	0.168	0.152	0.150	0.101	0.192	0.079	0.086

点击左侧目录树中的 graphs 图标，可见自动生成如下图形。

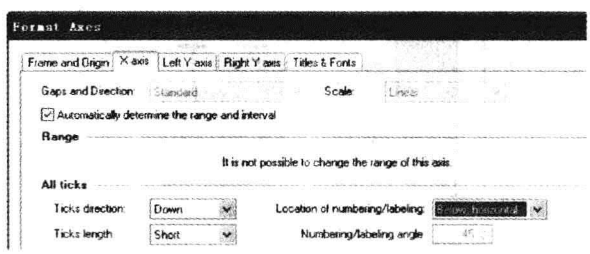


Step 3：**坐标轴和图形调整**

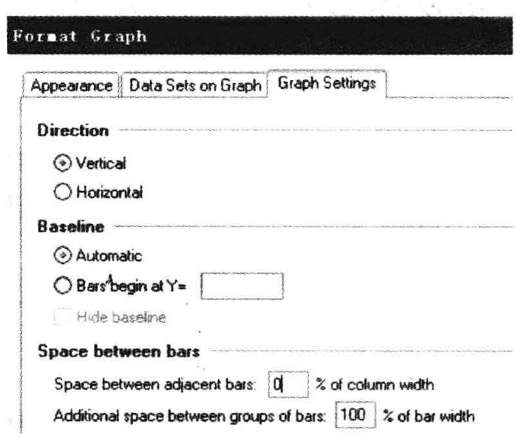
点击软件快捷菜单栏 Change 模块中的上左二，弹出坐标轴对话框。



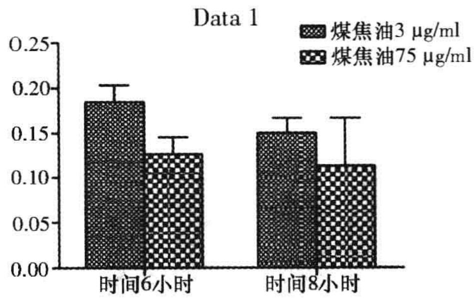
选择 X axis，将对 X 轴进行调整。在 All ticks 中对 Location of numbering/labeling 项进行选择，选中 Below，horizontal，即 X 轴刻度文字出现在轴下方，横向排列。



点击软件快捷菜单栏 change 模块中的上左三，弹出图形对话框。选择 Graph Settings，将对条图进行调整。

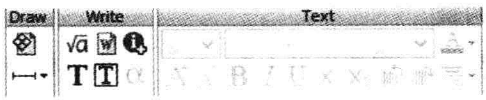


在 Space between bars 中 Space between adjacent bars：0% of column width，即每组内部条图之间的间隙为 0，默认为条图宽度的 50%。图形如下：

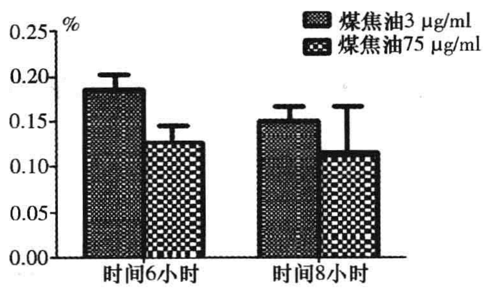


Step 4：**文字修饰**

将图形中不需要的文字点击删除即可。



点击快捷菜单栏 Write 中的 T(文字输入按钮)，然后在想要输入文字的地方输入文字即可，输入完成可以点击 Text 菜单栏中的选项进行字体大小等调整，最后图形如下：



四种不同处理情况下吸光度的值

第五节　重复测量资料方差分析的图形绘制

一、示例

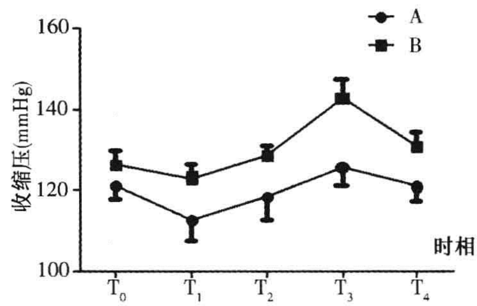
将手术要求基本相同的患者随机分成两组，在手术过程中分别采用 A，B 两种麻醉诱导方法，在 T_0(诱导前)、T_1、T_2、T_3、T_4 五个时相测量患者的收缩压，数据资料见下表。

不同麻醉、不同时相患者的收缩压(mmHg)

诱导方法	麻醉诱导时相				
	T_0	T_1	T_2	T_3	T_4
A	121.00 ± 3.54	112.40 ± 5.13	118.40 ± 5.64	125.80 ± 4.71	120.80 ± 3.70
B	126.20 ± 3.63	123.00 ± 3.39	128.60 ± 1.95	142.60 ± 4.83	130.60 ± 3.71

二、该示例的图形表达

本图形不能通过原始数据直接得出,但是数据表格可以通过 SPSS 重复测量的方差分析结果中得到。图形称之为误差线图,每个点表示每个单元格的均数,误差线表示单元格的标准差,然后用直线将均数连接起来,表达一种趋势。



不同麻醉诱导不同时相患者的收缩压

三、图形绘制

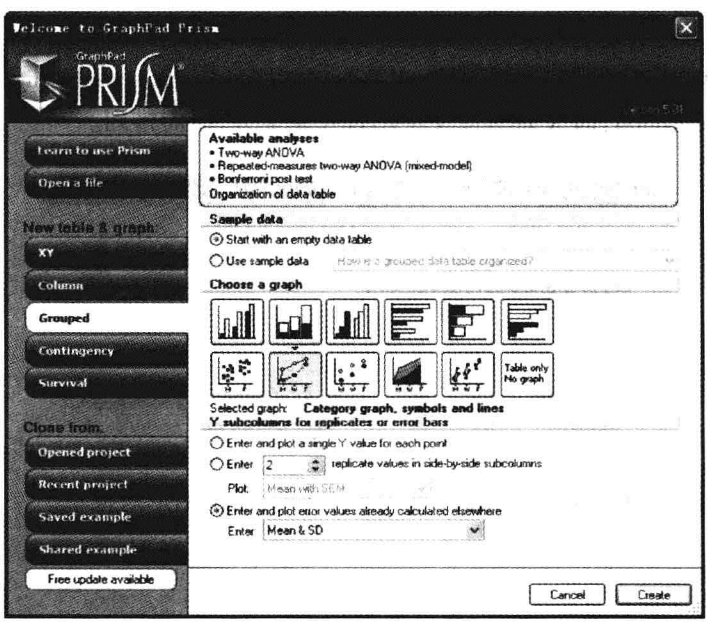
Step 1:图形类型选择

1. 左侧 New table & graph 项选择 Grouped,该类型适合两因素方差分析。

2. 右上 Sample data 项选择 Start with an empty data table,即建立新的数据文件。

3. 右中 Choose a graph 中选择左上的 Category graph, symbols and lines 图形类型。

4. 右下 Y subcolumns for replicates or error bars 中选择 Enter and plot error values already calculated elsewhere,即已经计算出了各单元格的均数和标准差;Plot 选择 Mean & SD,即表示均数和标准差。

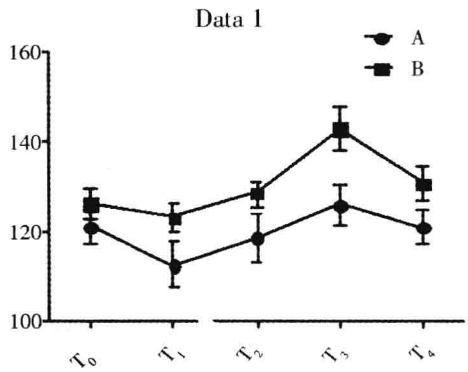


Step 2:**数据录入**

软件自动进入数据录入窗口,即 data 窗口,按下图录入数据。

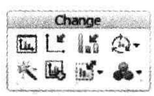
Table format Grouped	A		B	
	A		B	
	Mean	SD	Mean	SD
1 T0	121.00	3.54	126.20	3.63
2 T1	112.40	5.13	123.00	3.39
3 T2	118.40	5.64	128.60	1.95
4 T3	125.80	4.71	142.60	4.83
5 T4	120.80	3.70	130.60	3.71

点击左侧目录树中的 graphs 图标,可见自动生成如下图形。

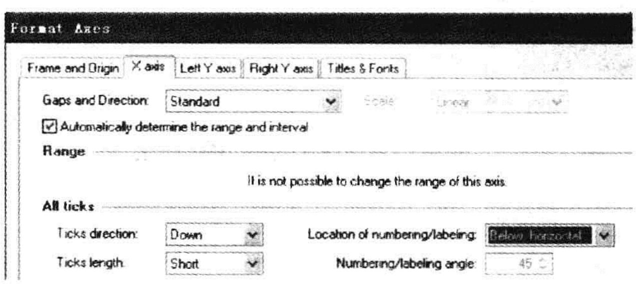


Step 3:**坐标轴和图形调整**

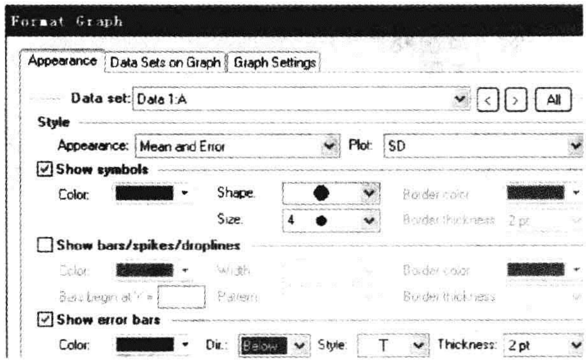
点击软件快捷菜单栏 change 模块中的上左二,弹出坐标轴对话框。



选择 X axis,将对 X 轴进行调整。在 All ticks 中对 Location of numbering/labeling 项进行选择,选中 Below, horizontal,即 X 轴刻度文字出现在轴下方,横向排列。

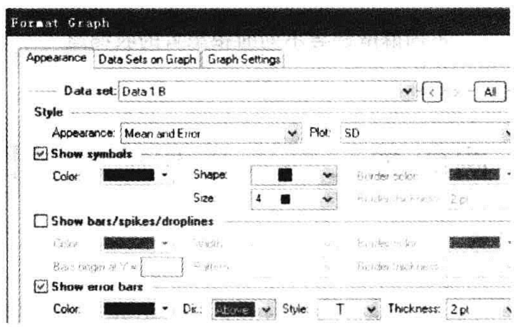


点击软件快捷菜单栏 change 模块中的上左三,弹出图形对话框。



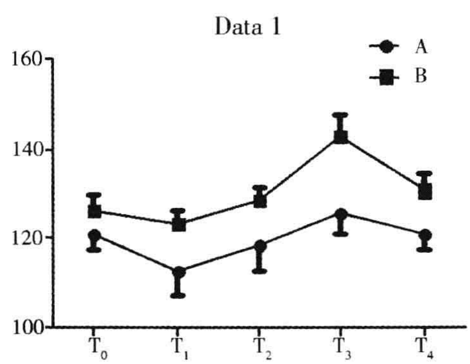
①Data set 项中选择 Data 1:A,即对 A 组数据对应的图形进行调整。

②Show error bars 项目中的 Dir(direction,方向),默认为 both,即同时显示上下误差线,此处改成 Below,只显示下误差线。



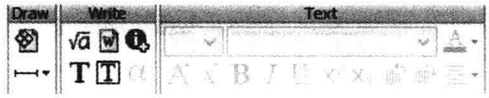
①Data set 项中选择 Data 1:B,即对 B 组数据对应的图形进行调整。

②Show error bars 项目中的 dir(direction,方向),默认为 both,即同时显示上下误差线,此处改成 Above,只显示上误差线。由于有两条线存在,同时显示上下误差线使图形不清晰,修改后一条误差线向上,而另一条误差线向下则使图形更加明了,至于各组数据的误差线方向,则需要根据具体数值来决定。结果图形如下:

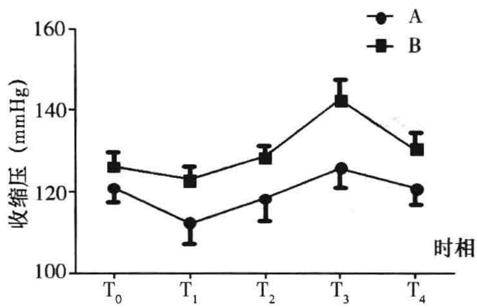


Step4:文字修饰

将图形中不需要的文字点击删除即可。



点击快捷菜单栏 Write 中的 T(文字输入按钮),然后在想要输入文字的地方输入文字即可,输入完成可以点击 Text 菜单栏中的选项进行字体大小等调整,最后图形如下:



不同麻醉诱导不同时相患者的收缩压

第六节　两组独立样本秩和检验的图形绘制

一、示例

用两种药物杀灭钉螺,每批用 200~300 只活钉螺,用药后清点钉螺的死亡数,并计算死亡率(%),结果见表,问两种药物杀死钉螺的效果有无差别。(《卫生统计学》第六版例 10-3,P181)

两种药物杀灭钉螺死亡率(%)的比较

甲药死亡率(%)	乙药死亡率(%)	甲药死亡率(%)	乙药死亡率(%)
32.5	16.0	49.0	32.5
35.5	22.5	49.5	38.0
40.5	26.0	51.5	40.5
40.5	28.5		

二、该示例的图形表达

1. 散点图(图 1)　清晰地描述了各点数据的情况,容易发现离群值,与成组 t 检验不同的是,散点中的横线为中位数(median)。当数据较少时散点图表达比较合适。

2. 箱式图(图 2)　由 Q_1(25%百分位数)和 Q_3(75%百分位数)构成箱式图的“箱体”部分,箱中的横线表示中位数(median),触须表示最大值和最小值。值得注意的是,触须也可以表示

10% ~90%,或者5% ~95%数值,在软件中可以选择。当数据较多时箱式图表达比较合适。

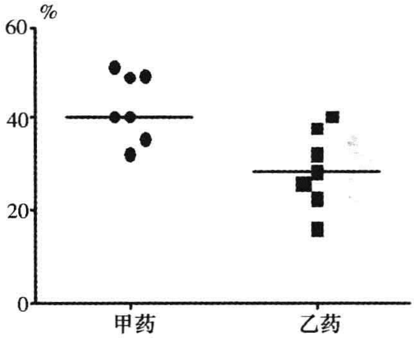


图 1　两种药物杀灭钉螺死亡率比较

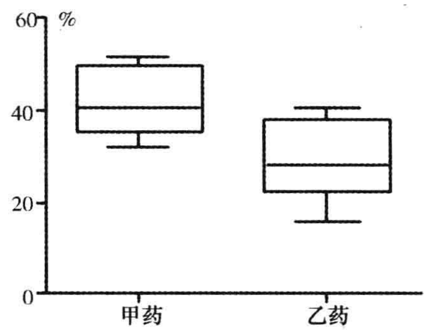


图 2　两种药物杀灭钉螺死亡率比较

三、图形绘制

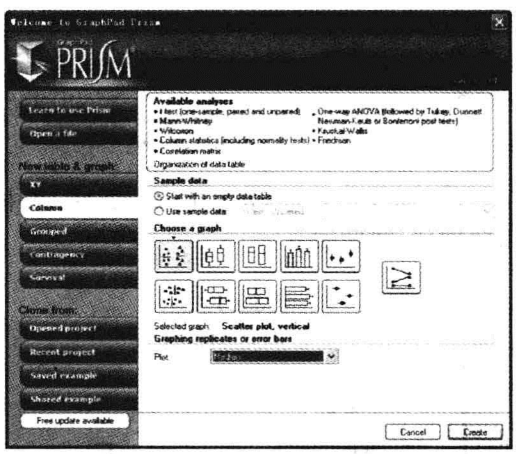
Step 1:图形类型选择

1. 左侧 New table & graph 项选择 Column,该类型适合常见的 t 检验、单因素方差分析和秩和检验。

2. 右上 Sample data 项选择 Start with an empty data table,即建立新的数据文件。

3. 右中 Choose a graph 中选择左上的 Scatter plot, vertical 图形类型,即纵向散点图。

4. 右下 Graphing replicates or error bars 中选择 Median,即用横线表示均数。

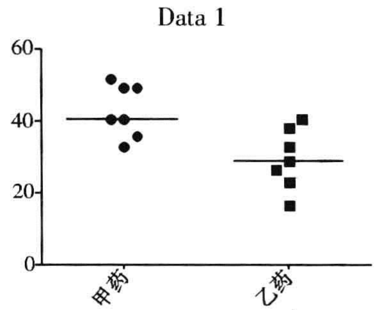


Step 2:数据录入

软件自动进入数据录入窗口,即 data 窗口,按下图录入数据。

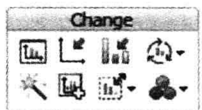
	A 甲药	B 乙药
	Y	Y
1	32.5	16.0
2	35.5	22.5
3	40.5	26.0
4	40.5	28.5
5	49.0	32.5
6	49.5	38.0
7	51.5	40.5

点击左侧目录树中的 graphs 图标,可见自动生成如下图形。

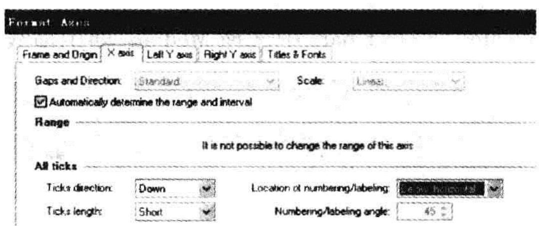


Step 3:**坐标轴和图形调整**

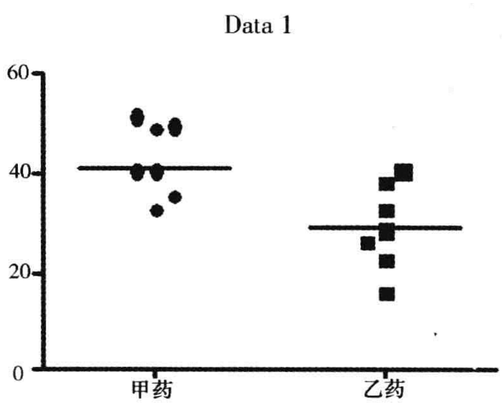
点击软件快捷菜单栏 change 模块中的上左二,弹出坐标轴对话框。



选择 X axis,将对 X 轴进行调整。

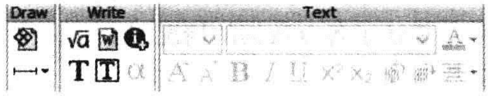


在 All ticks 中对 Location of numbering/labeling 项进行选择,选中 Below, horizontal,即 X 轴刻度文字出现在轴下方,横向排列,图形如下:

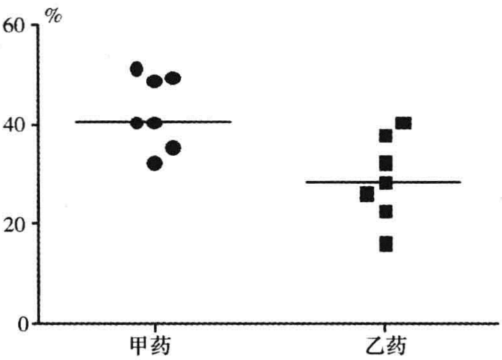


Step 4:**文字修饰**

将图形中不需要的文字点击删除即可。



点击快捷菜单栏 Write 中的 T(文字输入按钮),然后在想要输入文字的地方输入文字即可,输入完成可以点击 Text 菜单栏中的选项进行字体大小等调整,最后图形如下:



两种药物杀灭钉螺死亡率比较

第七节　多组独立样本秩和检验的图形绘制

一、示例

某医院用3种不同方法治疗15例胰腺癌患者,每种方法各治疗5例,治疗后生存月数见下表,问这3中方法对胰腺癌患者的疗效有无差别。(《卫生统计学》第六版例10-5,P183)

3种方法治疗胰腺癌患者的生存月数比较

甲法	乙法	丙法
3	6	2
4	9	3
7	10	5
8	12	7
8	13	8

二、该示例的图形表达

1. 箱式图(图1)　由 Q_1(25%百分位数)和 Q_3(75%百分位数)构成箱式图的"箱体"部分,箱中的横线表示中位数(median),触须表示最大值和最小值。值得注意的是,触须也可以

表示 10% ~90%，或者 5% ~95% 数值。当各组数据较多时箱式图表达比较合适。

2. 散点图(图 2)　清晰地描述了各点数据的情况，与成组 t 检验不同的是，散点中的横线为中位数(median)。当各组数据较少时散点图表达比较合适。

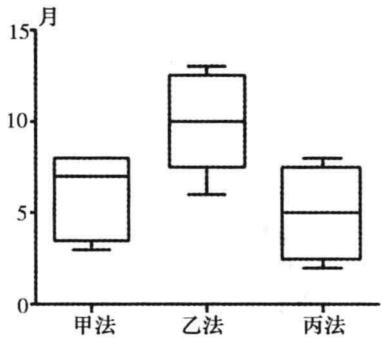


图 1　三种方法生存月数比较

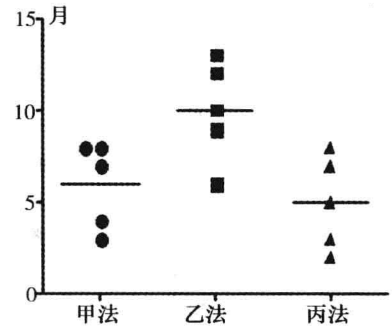


图 2　三种方法生存月数比较

三、图形绘制

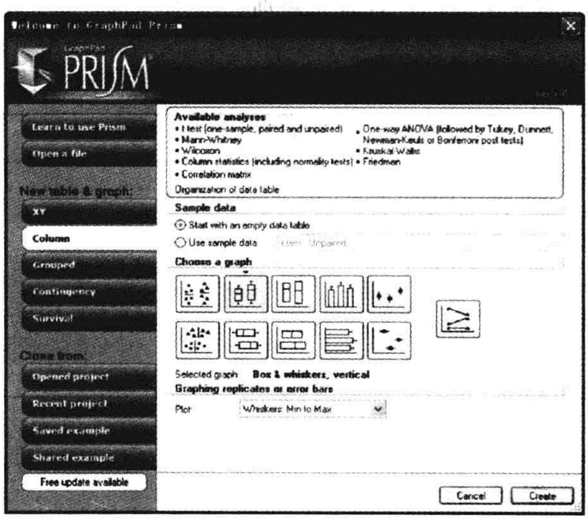
Step 1:**图形类型选择**

1. 左侧 New table & graph 项选择 Column，该类型适合常见的 t 检验、单因素方差分析和秩和检验。

2. 右上 Sample data 项选择 Start with an empty data table，即建立新的数据文件。

3. 右中 Choose a graph 中选择左上的 Box & whiskers, vertical 图形类型，即纵向箱式图。

4. 右下 Graphing replicates or error bars 中选择 Whiskers: Min to Max，即触须为最小值到最大值。

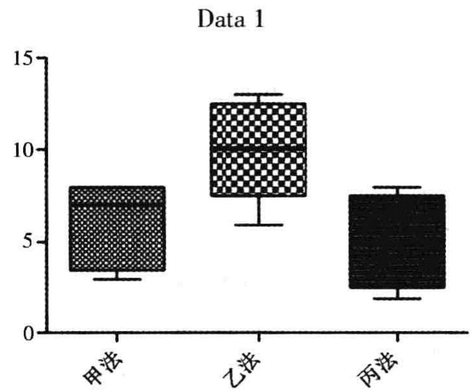


Step 2:**数据录入**

软件自动进入数据录入窗口，即 data 窗口，按下图录入数据。

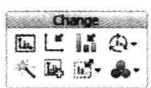
	A	B	C
	甲法	乙法	丙法
	Y	Y	Y
1	3	6	2
2	4	9	3
3	7	10	5
4	8	12	7
5	8	13	8

点击左侧目录树中的 graphs 图标,可见自动生成如下图形。

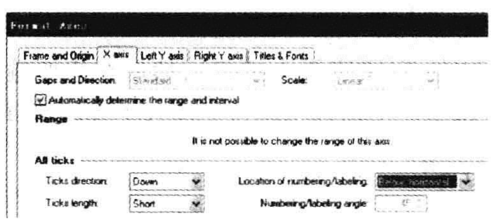


Step 3:**坐标轴和图形调整**

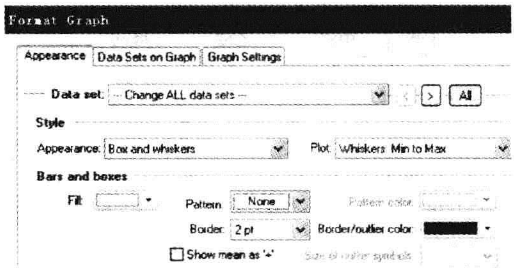
点击软件快捷菜单栏 change 模块中的上左二,弹出坐标轴对话框。



选择 X axis,将对 X 轴进行调整。在 All ticks 中对 Location of numbering/labeling 项进行选择,选中 Below, horizontal,即 X 轴刻度文字出现在轴下方,横向排列。

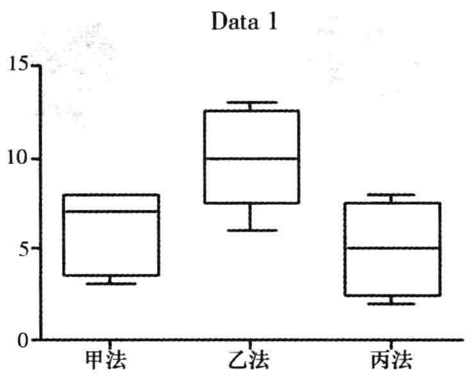


点击软件快捷菜单栏 change 模块中的上左三,弹出图形调整对话框。



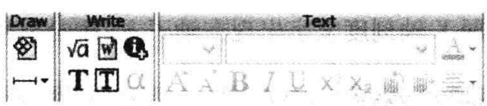
①Data set 项中选择 Change All data sets，即对所有数据集所表达的图形进行调整，当然也可以单独选择甲法、乙法或丙法进行调整。

②Bars and boxes 对箱式图进行调整，Fit 选择 clear，pattern 选择 none，即不对箱体进行填充，结果图形如下：

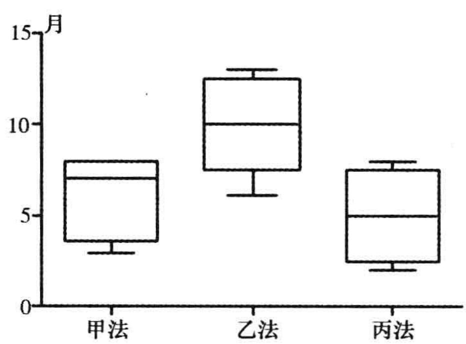


Step 4：**文字修饰**

将图形中不需要的文字点击删除即可。



点击快捷菜单栏 Write 中的 T(文字输入按钮)，然后在想要输入文字的地方输入文字即可，输入完成可以点击 Text 菜单栏中的选项进行字体大小等调整，最后图形如下：



第八节　简单线性回归和线性相关的图形绘制

一、示例

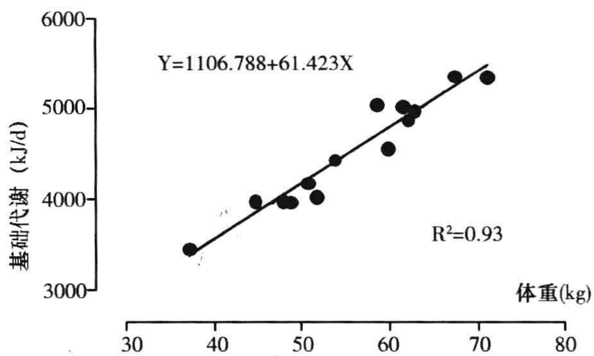
某地一项膳食调查中,随机抽取了 14 名 40~60 岁的健康妇女,测得每人的基础代谢(kJ/d)与体重(kg)数据,试分析这两项指标间有无关联。(《卫生统计学》第六版 例 11-1,P195)

14 名中年健康妇女的基础代谢与体重的测量值

编号	基础代谢(kJ/d)	体重(kg)	编号	基础代谢(kJ/d)	体重(kg)
1	4175.6	50.7	8	3970.6	48.6
2	4435.0	53.7	9	3983.2	44.6
3	3460.2	37.1	10	5050.1	58.6
4	4020.8	51.7	11	5355.5	71.0
5	3987.4	47.8	12	4560.6	59.7
6	4970.6	62.8	13	4874.4	62.1
7	5359.7	67.3	14	5029.2	61.5

二、该示例的图形表达

简单直线回归和线性相关一般是纠缠在一起的,在图上需要同时标注回归方程和决定系数(R^2)。



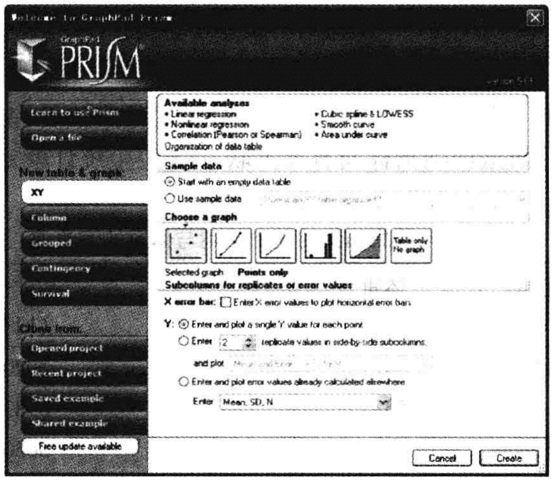
体重与基础代谢之间的关系

三、图形绘制

Step 1:图形类型选择

1. 左侧 New table & graph 项选择 XY,即线性回归分析。
2. 右上 Sample data 项选择 Start with an empty data table,即建立新的数据文件。

3. 右中 Choose a graph 中选择左上的 points only 图形类型，即散点图。
4. 右下 Y 选择 Enter and plot a single Y value for each point，即一个 X 对应一个 Y 值。

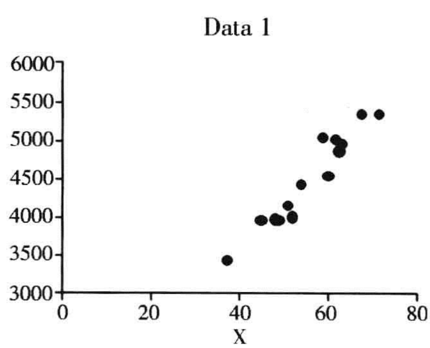


Step 2：**数据录入**

软件自动进入数据录入窗口，即 data 窗口，按下图录入数据。

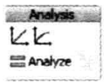
Table format: XY		X	A
		X	Y
		X	Y
1	Title	50.7	4175.6
2	Title	53.7	4435.0
3	Title	37.1	3460.2
4	Title	51.7	4020.8
5	Title	47.8	3987.4
6	Title	62.8	4970.6
7	Title	67.3	5359.7
8	Title	48.6	3970.6
9	Title	44.6	3983.2
10	Title	58.6	5050.1
11	Title	71.0	5355.5
12	Title	59.7	4560.6
13	Title	62.1	4874.4
14	Title	61.5	5029.2

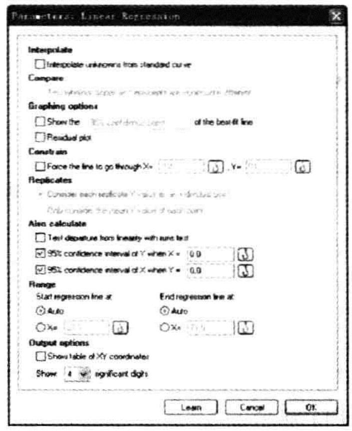
点击左侧目录树中的 graphs 图标，可见自动生成如下图形。



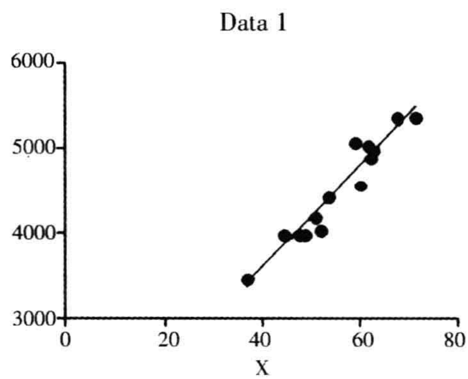
Step 3:**添加拟合直线**

点击软件菜单栏的 Analysis 模块中左上的直线回归对话框,弹出直线回归参数设定对话框(parameters: Linear Regression)。



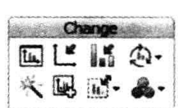


不需要进行任何设定,直接按 ok 即可,软件自动为图形添加拟合的直线,见下图:

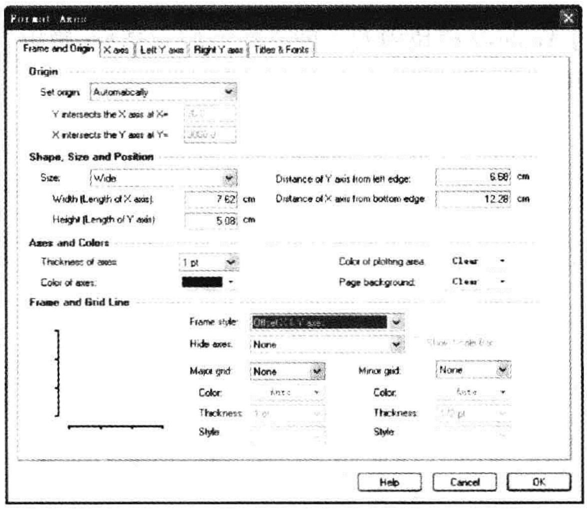


Step 4:**坐标轴和图形调整**

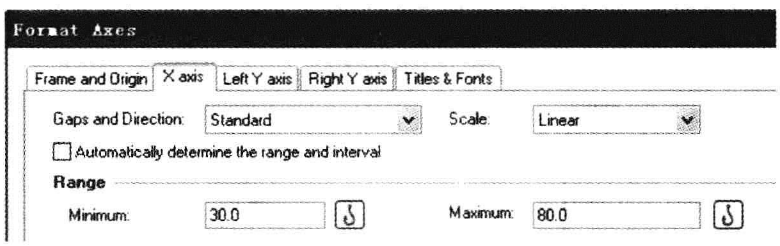
点击软件快捷菜单栏 change 模块中的上左二,弹出坐标轴对话框。



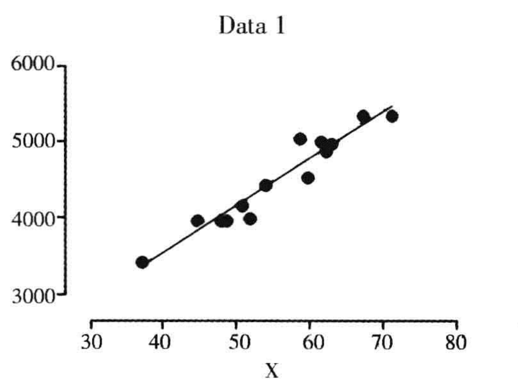
选择 Frame and Origin,将对坐标系进行整体调整。下方的 Frame and Grid Line 对 Frame style 进行设定,选择 Offset X & Y axes,即 X 轴与 Y 轴无交叉,这是对坐标轴不交叉于(0,0)的一种处理方法。



选择 X axis,对 X 轴进行调整。

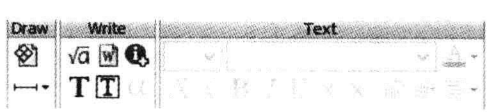


去除 Automatically determine the range and interval 前面方框中的选项,从而自定义 X 轴的刻度方法,在 Range 选项中设定 Minimum(最小值)为 30,Maximum(最大值)为 80,得到图形如下:

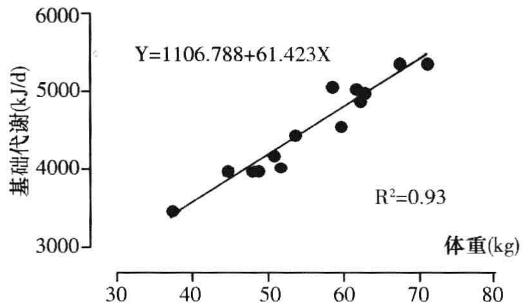


Step 5:**文字修饰**

将图形中不需要的文字点击删除即可。



点击快捷菜单栏 Write 中的 T(文字输入按钮),然后在想要输入文字的地方输入文字即可,输入完成可以点击 Text 菜单栏中的选项进行字体大小等调整,最后图形如下:



体重与基础代谢之间的关系

第九节　列联表分析的图形绘制

一、示例

某医师研究物理疗法、药物治疗和外用药膏三种疗法治疗周围性神经麻痹的疗效,资料见下表,问三种疗法的有效率有无差别。

三种疗法有效率的比较

疗法	有效	无效	合计	有效率(%)
物理治疗	68	54	122	56
药物治疗	75	44	119	63
外用治疗	50	55	105	48
合计	193	153	346	56

二、该示例的图形表达

列联表数据能用图形表达的比较少,常见的情况结局为等级资料的数据用条图和百分条图来表示,虽然在统计方法上这种数据可能采用秩和检验,但在图形表达上则不同于前面所描述的计量资料的秩和检验中用散点图和箱式图来表达。

1. 条图(图 1)　条图的高度反映了各组的具体例数(n)。

2. 百分条图(图 2)　条图的长度为 100%,而条图是由各结局组合而成的。该图形可以用扇形图来表达,但是 Graphpad Prism 不能绘制扇形图。

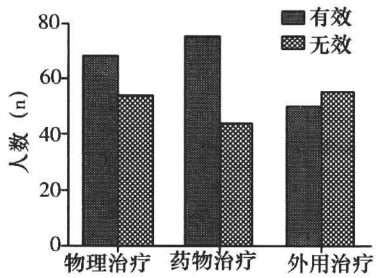


图 1　三种疗法疗效比较

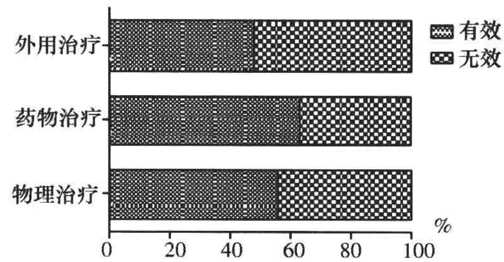


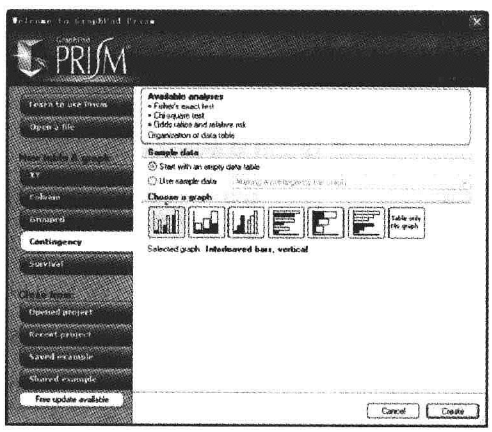
图 2　三种疗法有效率比较

三、图形绘制

(一)条图的绘制

Step 1:图形类型选择

1. 左侧 New table & graph 项选择 Contingency,即常说的列联表分析。
2. 右上 Sample data 项选择 start with an empty data table,即建立新的数据文件。
3. 右下 Choose a graph 中选择左侧的 Interleaved bars, vertical 图形类型,即分组条图。

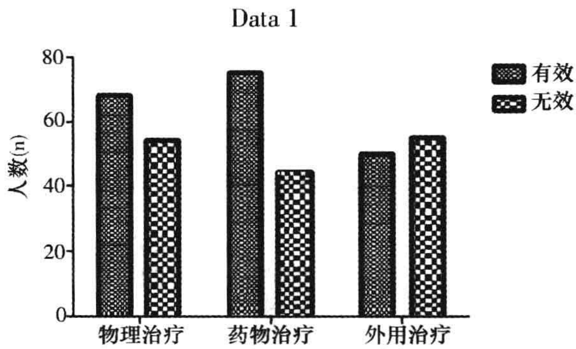


Step 2:数据录入

软件自动进入数据录入窗口,即 data 窗口,按下图录入数据。

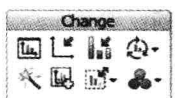
Table format: Contingency	A	B
	有效	无效
	Y	Y
1 物理治疗	68	54
2 药物治疗	75	44
3 外用治疗	50	55

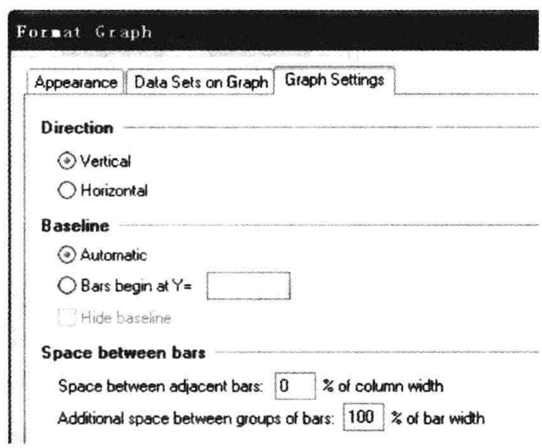
点击左侧目录树中的 graphs 图标,可见自动生成如下图形。



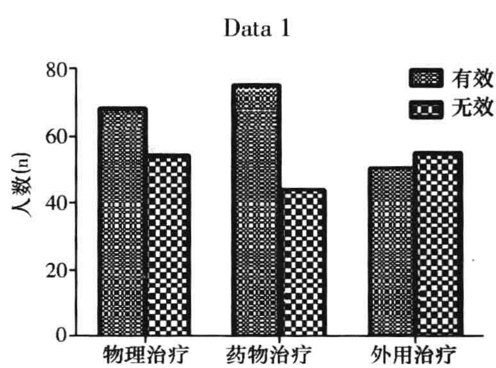
Step 3:**坐标轴和图形调整**

点击软件快捷菜单栏 change 模块中的上左三，弹出图形对话框。



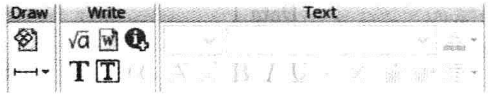


选择 Graph Settings,对图形整体进行调整。在 Space between bars 中对组内条图之间的距离进行设定:Space between adjacent bars:0% of column width,组内条图距离为 0,图形如下:

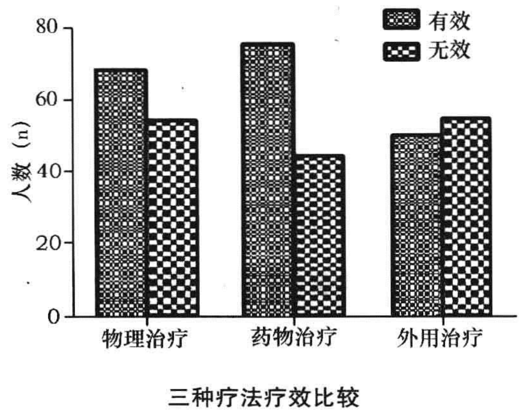


Step4:**文字修饰**

将图形中不需要的文字点击删除即可。



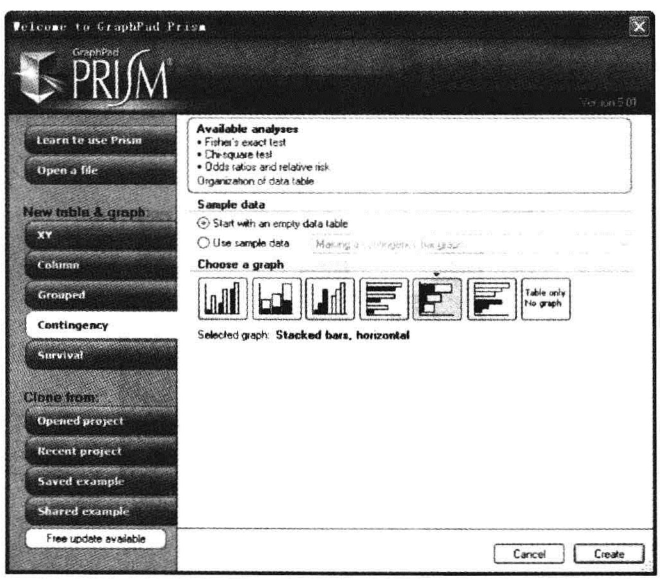
点击快捷菜单栏 Write 中的 T(文字输入按钮),然后在想要输入文字的地方输入文字即可,输入完成可以点击 Text 菜单栏中的选项进行字体大小等调整,最后图形如下:



(二)百分条图的绘制

Step 1:图形类型选择

1. 左侧 New table & graph 项选择 Contingency,即常说的列联表分析。
2. 右上 Sample data 项选择 start with an empty data table,即建立新的数据文件。
3. 右下 Choose a graph 中选择左侧的 Stacked bars, horizontal 图形类型,即累积条图。



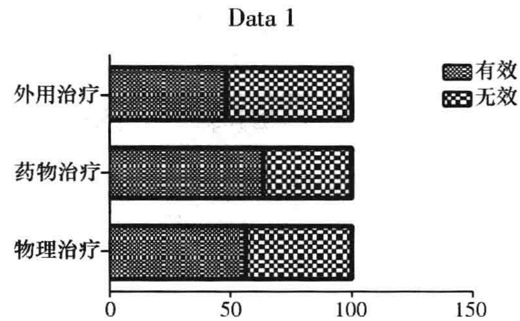
Step 2:数据录入

软件自动进入数据录入窗口,即 data 窗口,特别注意,由于软件不能直接作出百分条图,

我们需要在数据窗口输入百分数,而非原始例数,按下图录入数据。

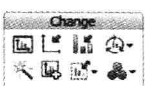
Table format: Contingency		A 有效	B 无效
		Y	Y
1	物理治疗	56	44
2	药物治疗	63	37
3	外用治疗	48	52

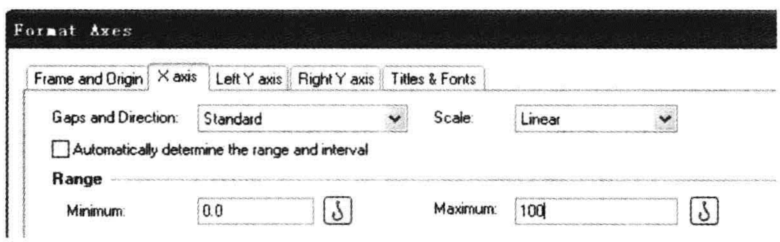
点击左侧目录树中的 graphs 图标,可见自动生成如下图形。



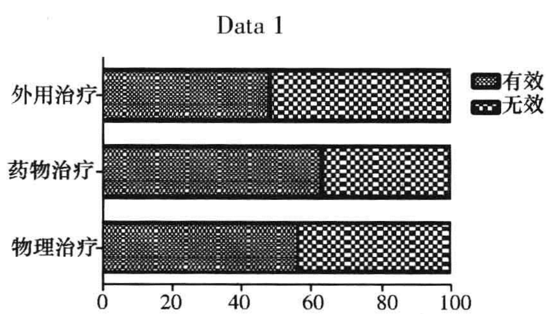
Step 3:**坐标轴和图形调整**

点击软件快捷菜单栏 change 模块中的上左二,弹出坐标轴对话框。



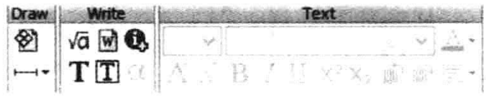


选择 X axis,将对 X 轴进行调整。去除 Automatically determine the range and interval 前面方框中的选项,从而自定义 X 轴的刻度,在 Range 选项中设定 Minimum(最小值)为 0,Maximum(最大值)为 100,得到图形如下:

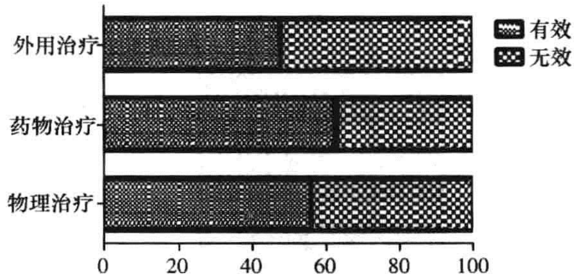


Step4:**文字修饰**

将图形中不需要的文字点击删除即可。



点击快捷菜单栏 Write 中的 T(文字输入按钮),然后在想要输入文字的地方输入文字即可,输入完成可以点击 Text 菜单栏中的选项进行字体大小等调整,最后图形如下:



第十节　生存分析的图形绘制

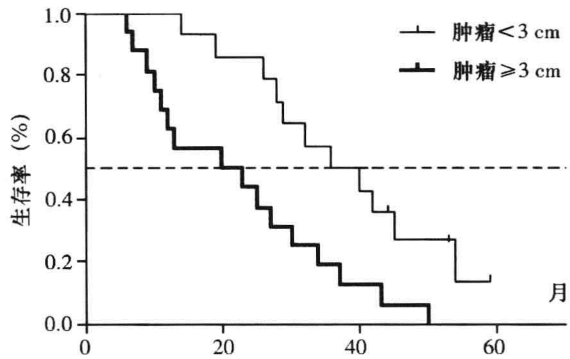
一、示例

14 例膀胱肿瘤 <3.0 cm 患者和 16 例膀胱肿瘤 ≥3.0 cm 患者的生存时间(月)如下,试估计两组生存率。(《卫生统计学》第六版例 21 -2,P403)

肿瘤 < 3.0 cm	14	19	26	28	29	32	36	40	42	44⁺
	45	53⁺	54	59⁺						
肿瘤 ≥3.0 cm	6	7	9	10	11	12	13	20	23	25
	27	30	34	37	43	50				

二、该示例的图形表达

1. 生存曲线一般成阶梯状,每出现 1 例死亡,阶梯就下降一阶。
2. 删失数据用"|"线表示。
3. 中位生存时间(median survival time)表示刚好有 50% 的个体其存活期大于该时间,即从生存率 Y =0.5 处划出一条参考线,该线与各组生存曲线的交叉,其交点的 X 轴刻度为该组的中位生存时间。如该图中肿瘤 <3 cm 患者的中位生存时间为 38 月,肿瘤 ≥3 cm 患者的中位生存时间为 20 月。



两组肿瘤患者生存曲线的比较

三、图形绘制

Step 1:**图形类型选择**

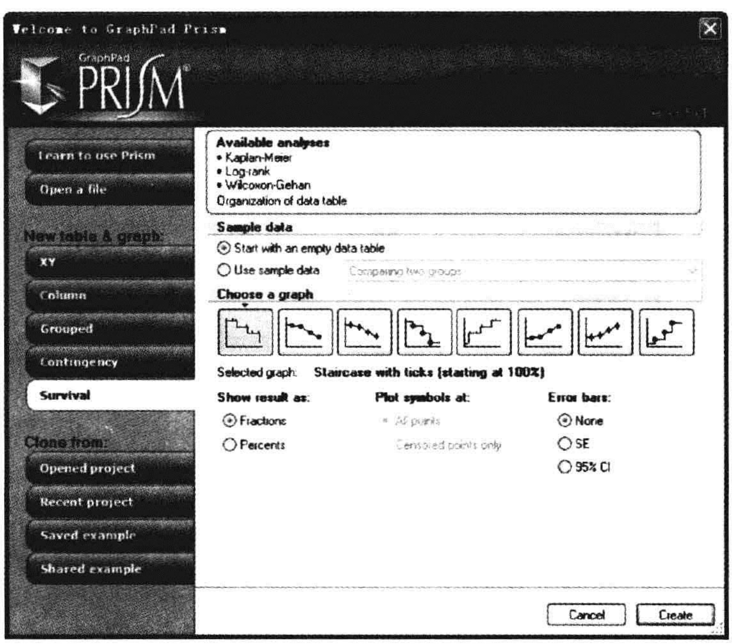
1. 左侧 New table & graph 项选择 Survival,即生存分析。

2. 右上 Sample data 项选择 Start with an empty data table,即建立新的数据文件。

3. 右中 Choose a graph 项选择 Staircase with ticks (starting at 100%),即建立阶梯型生存曲线,起点为 100%,删失值用"|"表示。。

4. Show result at 选择 Fractions,表示 Y 轴累计生存率为 0 到 1,若 Percents,则表示 Y 轴累计生存率 0 ~ 100%,两者的含义一样,选择纯属个人喜好,本次选择 Fractions。

5. Error bars 设定是否显示误差线,误差线能带来更多的信息,但是使图形显得过分拥挤,一般选择不显示(None)。

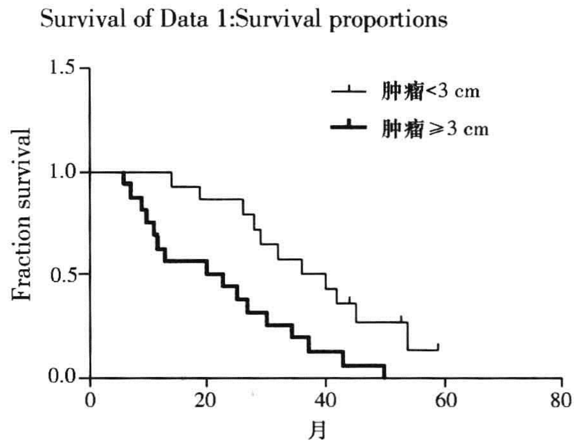


Step 2:**数据录入**

软件自动进入数据录入窗口,即 data 窗口,按下图录入数据。X 纵轴表示生存时间,各组数据分别占一列,如此处的"肿瘤 <3 cm"和"肿瘤 ≥3 cm"。X 纵列若出现相同的值,重复输入即可。Y 纵列完全数据用"1"表示,删失数据用"0"表示。

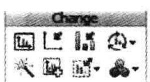
Table format: **Survival**		X	A	B
		月	肿瘤<3cm	肿瘤≥3cm
		X	Y	Y
1	Title	14	1	
2	Title	19	1	
3	Title	26	1	
4	Title	28	1	
5	Title	29	1	
6	Title	32	1	
7	Title	36	1	
8	Title	40	1	
9	Title	42	1	
10	Title	44	0	
11	Title	45	1	
12	Title	53	0	
13	Title	54	1	
14	Title	59	0	
15	Title	6		1
16	Title	7		1
17	Title	9		1
18	Title	10		1
19	Title	11		1
20	Title	12		1
21	Title	13		1
22	Title	20		1
23	Title	23		1
24	Title	25		1
25	Title	27		1
26	Title	30		1
27	Title	34		1
28	Title	37		1
29	Title	43		1
30	Title	50		1

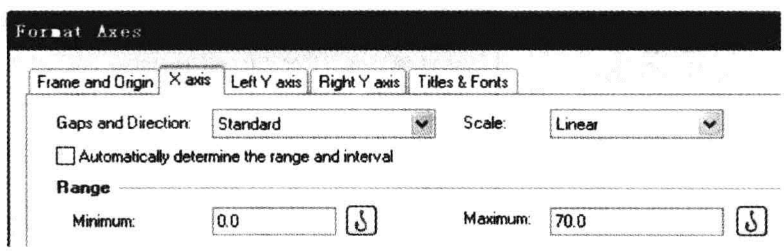
点击左侧目录树中的 graphs 图标,可见自动生成如下图形。



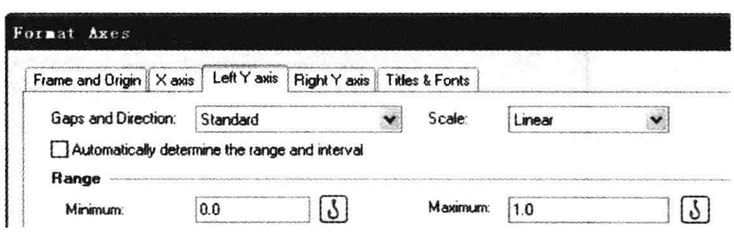
Step 3:**坐标轴调整**

点击软件快捷菜单栏 change 模块中的上左二，弹出坐标轴对话框。选择 X axis，对 X 轴进行调整。去掉 Automatically determine the range and interval，对 Range 进行调整，Minimum 设为 0，Maximum 设为 70。

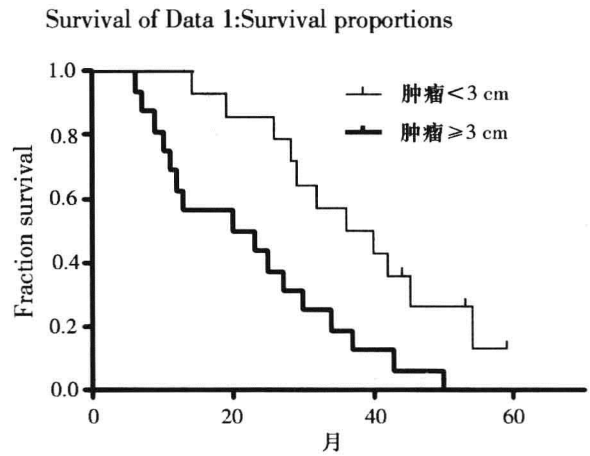




选择 Left Y axis，将对 Y 轴进行调整。

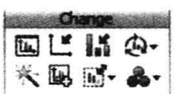


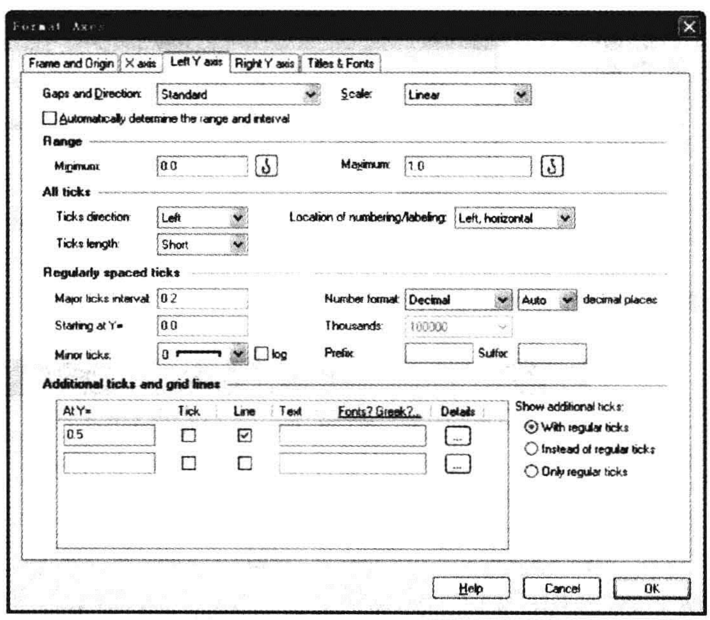
去掉 Automatically determine the range and interval，对 Range 进行调整，Minimum 设为 0，Maximum 设为 1，图形如下：



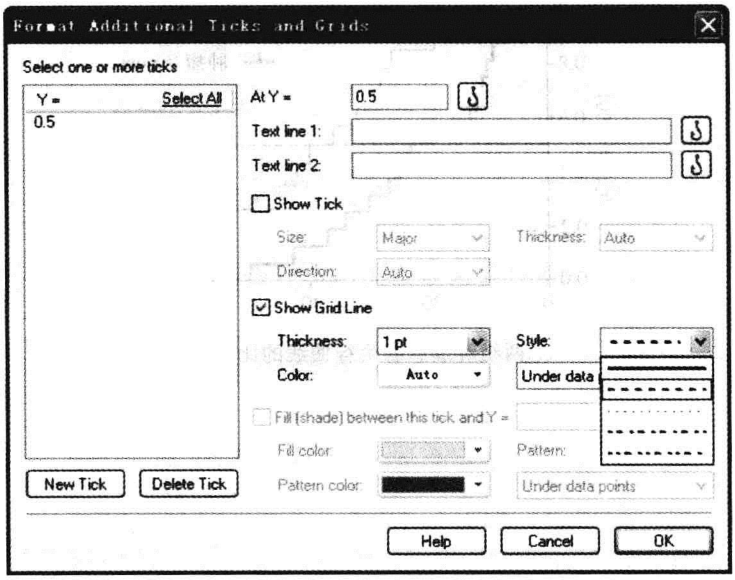
Step4:**添加中位生存参考线**

点击软件快捷菜单栏 change 模块中的上左二，弹出坐标轴对话框。选择 Left Y axis，将添加中位生存参考线。在 Additional ticks and grid lines 下方，填入 At Y＝0.5，选中 Line 对话框，即在 Y＝0.5 处添加一条参考线。点击 Details 按钮进入参考线设置。

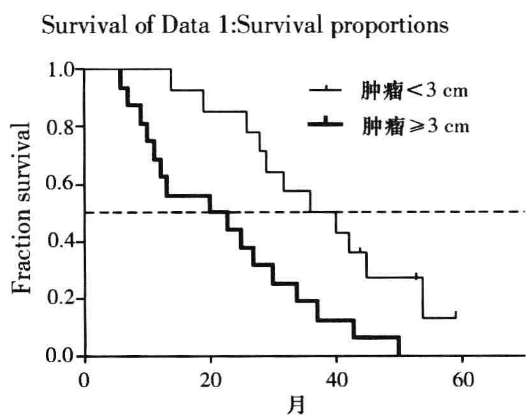




进入 Format Additional Ticks and Grids，对参考线进行设置。

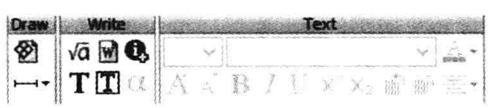


Show Grid Line 下方 Style 选择“粗虚线”。形成图形如下：

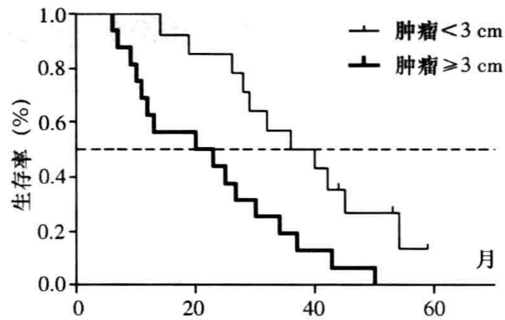


Step 4:**文字修饰**

将图形中不需要的文字点击删除即可。



点击快捷菜单栏 Write 中的 T(文字输入按钮)，然后在想要输入文字的地方输入文字即可，输入完成可以点击 Text 菜单栏中的选项进行字体大小等调整，最后图形如下：



两组肿瘤患者生存曲线的比较

>> 第十七章　统计图的排版与导出

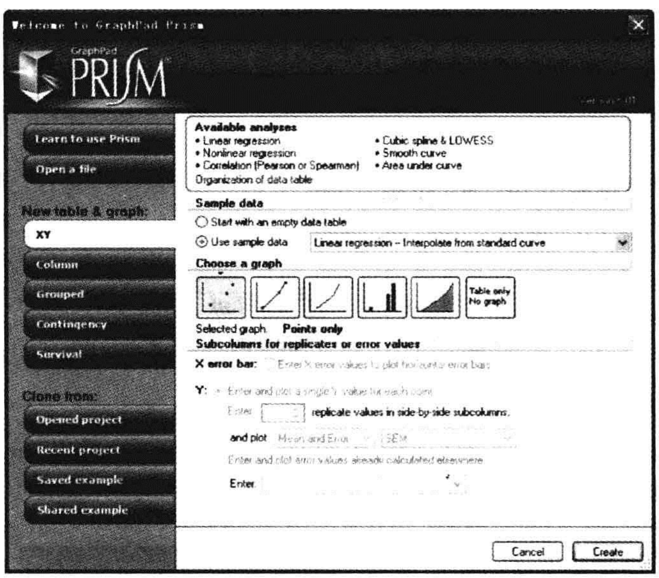
一、建立示例文件

Step1:利用软件自带示例数据建立文件“排版1. pzf”

1. 在 New table & graph 中选择 XY。

2. 在 Sample data 中选择 Use sample data,即采用软件自带数据库,从中选择 Linear regression—Interpolate from standard curve。

3. 点击 create 按钮,自动生成数据和图形。



4. 将图形保存为“排版1. pzf”,关闭软件。

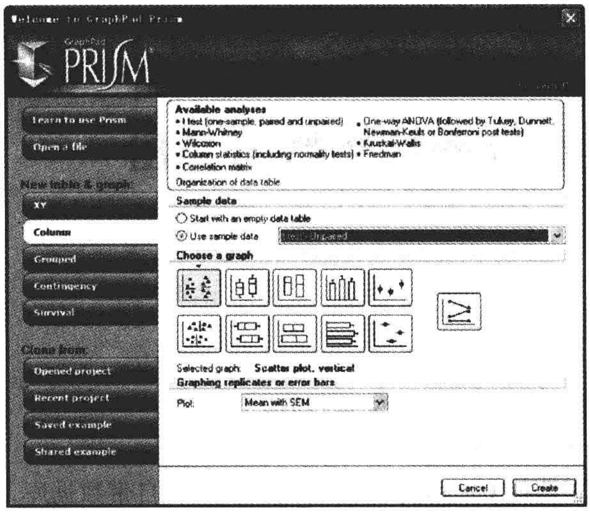


Step 2:利用软件自带示例数据建立文件“排版2. pzf”

1. 在 New table & graph 中选择 Column

2. 在 Sample data 中选择 Use sample data,即采用软件自带数据库,从中选择 t test--Unpaired

3. 点击 create 按钮,自动生成数据和图形



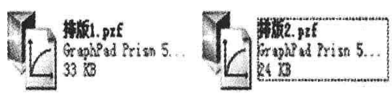
4. 将图形保存为“排版 2. pzf”,关闭软件



二、多图排版

Step 1:打开文件“排版 1. pzf”和“排版 2. pzf”

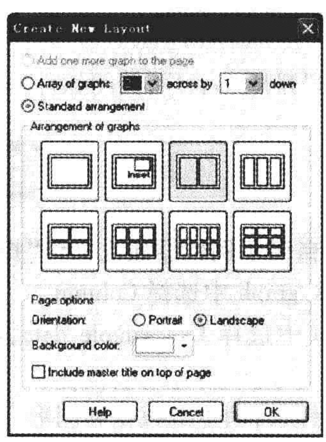
双击生成的文件,打开“排版 1. pzf”和“排版 2. pzf”



Step 2:设定版式

在打开“排版 1. pzf”文件中点击目录树中的 layouts,软件自动弹出 Create New Layout 对话框。该对话框是对多个图形进行排版的版式设定。





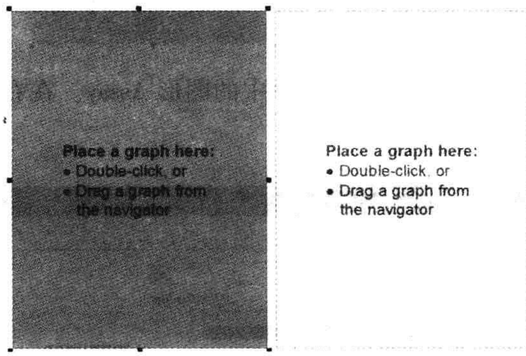
1. Standard arrangement　按照下面的缩略图,选用需要排版的样式,如果此处的样式不能满足你的要求,则可以从上方 Array of graphs 中自己设定。

2. Page options　设定版式的方向(Orientation),常见的为 portrait(纵向)和 landscape(横向),当版式方向发生改变时,上方的 arrangement of graphs 也发生相应的变化。

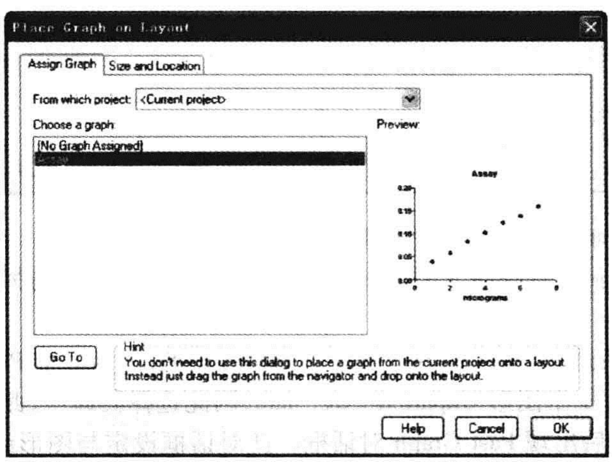
3. Background color　设定背景颜色,默认为透明,一般选择默认项即可。

Step 3:**将各图纳入版式中**

1. 各方框将选定需要纳入的图形。



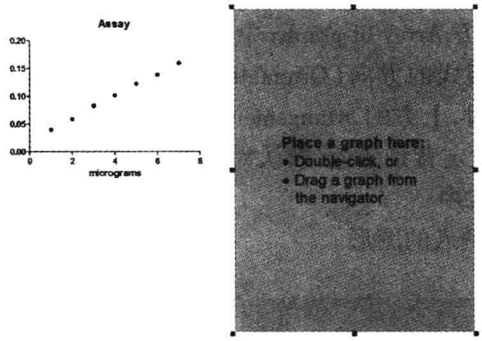
2. 在左侧方格上双击鼠标,自动跳出图形选择对话框(Place Graph on Layout)。



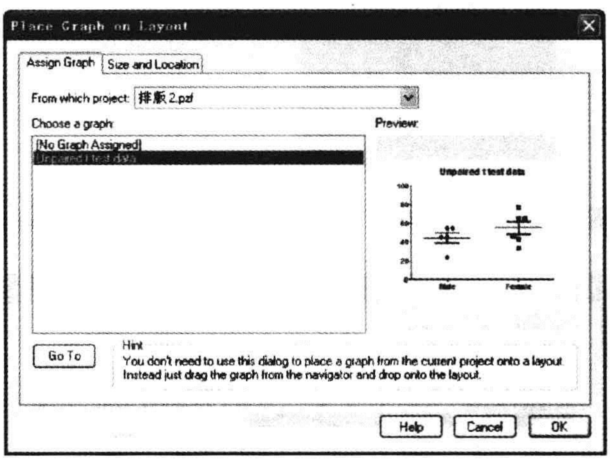
①选定 Assign Graph 选项,即选择图形。

②From which project:选择纳入图形的相应文件,此处保留默认选项 current project,即当前操作的文件"排版 1. pzf"。

③Choose a graph:选择"排版 1. pzf"文件中的具体图形,因为一个 PZF 文件可以包括多个图形,但本文件中只有一个图形 Assay,因此选择 Assay 即可,此时对话框右侧出现预览图(Preview)。确认后将图形 Assay 纳入版式中,见下图:



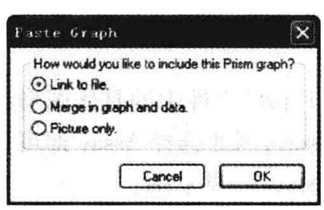
3. 左侧方格已经选入了“排版 1. pzf”文件中的图形 Assay。在右侧方格上双击鼠标，弹出图形设定对话框。



①选定 Assign Graph 选项，即选择图形。

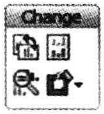
②From which project：选择纳入图形的相应文件，此处选择打开的另外一个文件“排版 2. pzf”。

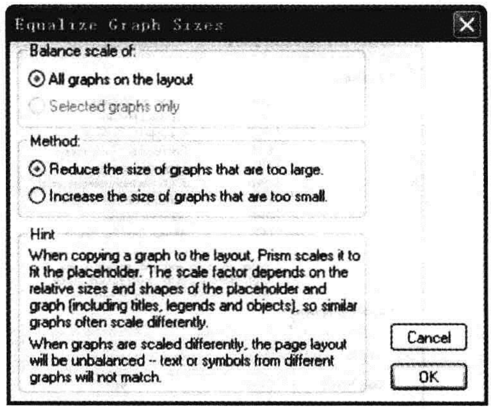
③Choose a graph：选择“排版 2. pzf”文件中的具体图形，因为一个 PZF 文件可以包括多个图形，但本文档中只有一个图形 Unpaired t test data，因此选择其即可，此时对话框右侧出现预览图（Preview）。确认后出现 Past Graph 对话框。此对话框设定与图形纳入排版的方式，默认为 Link to file（链接到文件），即该图形与源文件保持一致，保留默认选项即可。



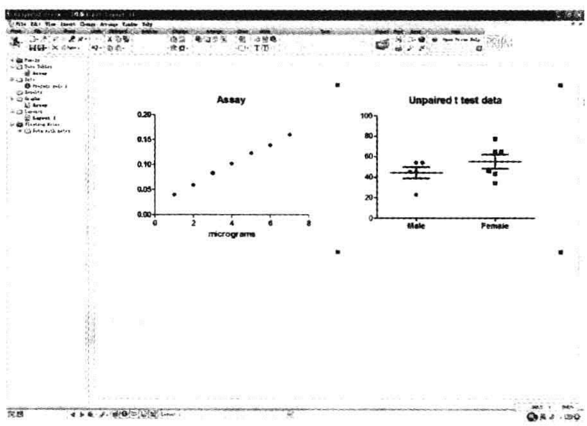
Step 4：**统一各图大小**

选择快捷菜单栏中右下的图形大小统一按钮，弹出 Equalize Graph Sizes 对话框。





Method 选择项中,Reduce the size of graphs that are too large,即以最小尺寸图形为标准进行调整;而 Increase the size of graphs that are too small,即以最大尺寸图形为标准进行调整,本例按默认项即可。排版完成后软件界面如下:

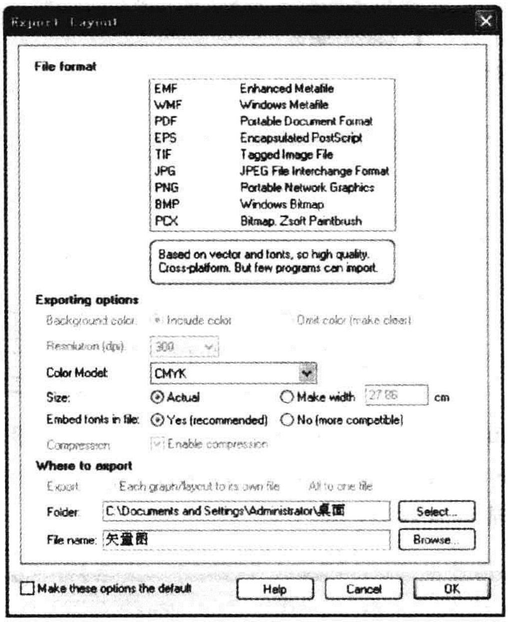


三、导出出版级图片

点击软件快捷菜单栏中的 Export 按钮,可以导出出版级图片。



（一）方法一：导出矢量图



1. File format　选择 EPS(Encapsulated PostScript)，即通行矢量图。

2. Exporting options　Code Model(色彩模式)，默认为 CMKY，即印刷模式。

3. Where to export　导出选项，Folder 文件存储位置，本次选择存放在桌面；File name 文件名，本次取名“矢量图”。

（二）方法二：导出位图



1. File format　选择 TIF(Tagged Image File)，即通行位图。

2. Exporting options　Resolution(dpi)，即选择分辨率，一般选择 600 即可；Code Model(色

彩模式),默认为CMKY,即印刷模式。

3. Where to export　导出选项,Folder文件存储位置,本次选择存放在桌面;File name文件名,本次取名“位图”。

四、注意事项

(一)矢量图和位图的区别

计算机绘图分为位图和矢量图两大类。

1. 位图图像也称为栅格图像,Photoshop及其他绘图软件一般都使用位图图像。位图图像由像素组成,每个像素都被分配一个特定位置和颜色值。在处理位图图像时,您编辑的是像素而不是对象或形状,也就是说,编辑的是每一个点。位图图像与分辨率(dpi)有关,即在一定面积的图像上包含有固定数量的像素。因此,如果在屏幕上以较大的倍数放大显示图像,或以过低的分辨率打印,位图图像会出现锯齿边缘。

2. 矢量图形由矢量定义的直线和曲线组成,Adobe Illustrator、CorelDraw、CAD等软件是以矢量图形为基础进行创作的。矢量图形根据轮廓的几何特性进行描述。图形的轮廓画出后,被放在特定位置并填充颜色。移动、缩放或更改颜色不会降低图形的品质。矢量图形与分辨率无关,可以将它缩放到任意大小和以任意分辨率在输出设备上打印出来,都不会影响清晰度。因此,矢量图形是文字(尤其是小字)和线条图形(比如徽标)的最佳选择。

(二)RGB模式和CMKY模式

1. RGB模式,自然界中绝大部分的可见光谱可以用红、绿和蓝三色光按不同比例和强度的混合来表示。RGB分别代表着3种颜色:R代表红色,G代表绿色、B代表蓝色。RGB模型也称为加色模型,RGB模型通常用于光照、视频和屏幕图像编辑。

2. CMYK色彩模式以打印油墨在纸张上的光线吸收特性为基础,图像中每个像素都是由靛青(C)、品红(M)、黄(Y)和黑(K)色按照不同的比例合成。在制作用于印刷色打印的图像时,要使用CMYK色彩模式。

(三)印刷要求和网络发表要求

1. *印刷要求*　按照杂志社的要求,可以采用矢量图(如EPS格式)和位图(如TIF格式),矢量图的大小很小,而位图可以高达几十兆,那么导出后可以Photoshop软件压缩成LZW格式,这种格式是无损压缩,图像质量不受到影响。由于GraphPad Prism软件对中文的兼容性不是太好,图形的中文标题排版不尽人意,可以导出后在制图软件(如Adobe Photoshop)当中修改即可。

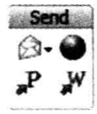
2. *网络发表要求*　现在很多杂志同时有纸质版和网络版两种,网络版的图片体积不能太大,多采用PNG格式和JPGE格式。

五、图片发布至WORD文档或PPT中

一键发布至word和幻灯片上。

1. 发布至word文档中,只需要点击快捷菜单Send模块中“W”按钮即可。

2. 发布至幻灯片文档中,只需要点击快捷菜单Send模块中“P”按钮即可。



六、注意事项

修改图形必须回到原图,而不能在排版后的图形中进行修改,而在幻灯片中,一般需要将图形变成彩图,因此也需要回到原图进行色彩配置,即选择快捷菜单栏中的 change 模块右下角中即可。

> > 附录一 EpiData,SPSS,GraphPad Prism 介绍

《论语·卫灵公》中有句名言"工欲善其事,必先利其器",含义为做好一件事情,准备工作非常重要。要进行数据的处理、分析和图示,选用适当的软件就至关重要,选择有时往往比努力更加重要。同时要撇弃那种想用一种软件完成所有任务的思想,如同一个大厨要烧制一顿丰盛的晚宴,仅仅一口好锅是远远不够的,至少需要勺、铲等器具。每种器具都有自己的优点,而不同的软件也具有自身的特点,因此本书采用 EpiData 进行数据录入、SPSS 进行统计分析和 GraphPad Prism 绘制图形。我们不要做软件的奴隶,而要做软件的主人,让各种软件为我所用,联合使用,又快又好地完成数据处理。

一、EpiData 简介

(一)选择 EpiData 的理由

大家可能有这样的疑问:"数据录入还需要软件,不是用 Excel 或者 SPSS 自带的数据录入窗口进行数据录入不就可以了吗?"当然,如果数据量少,用 Excel 或 SPSS 录入数据就足够了,但如果是大规模的流行病学调查,调查表就有上十页,问题多达几十个,这时候用 Excel 或 SPSS 就有点勉强了,因此出现了专门的数据录入软件,如微软 OFFICE 组件中的 ACCESS, SPSS 公司的 Data Entry。不过此处隆重推荐 EpiData,其入围理由如下:

1. 免费,有句名言"世界上没有免费的午餐",这句话放在这里就不对了,数据录入软件有个免费的午餐 EpiData,有这样的美事,大家还不赶快享用。

2. 全中文界面,通俗易懂。

3. 与调查表格式非常近似,为了录入的方便,最好的录入软件界面应当尽量保持与纸质调查表一致的效果,这样才能保证录入的速度和质量,而 EpiData 却能够做到这点。

4. 双录入功能避免录入错误,双录入是数据录入中质量控制的一个重要环节。

5. 丰富的数据导出格式,可以导出成世界流行的统计分析软件 SAS、SPSS 和 Stata 等数据格式,实现从数据录入到统计分析的无缝链接。

(二)EpiData 的历史

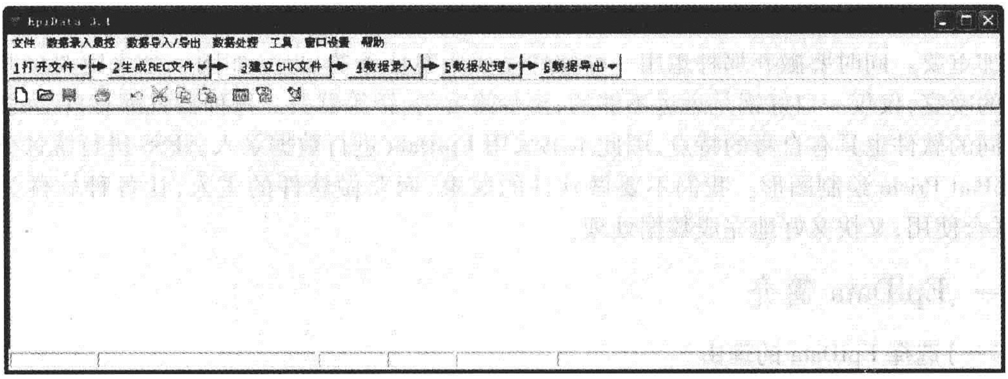
作为"事故预防项目"的一个部分,芬兰的 Jens M. Lauritsen 博士开始设想开发 EpiData,其开发动机来自于当时 EpiInfo 6 不能在 Dos 环境中运行,而当时市场上的商业数据软件没有一款专注于数据记录、使用简单和双录入验证。Mark Myatt 和程序员 Michael Bruus 也先后加入了 EpiData 开发团队,其目的在于开发一个简单易用的独立程序,并且该软件不需要其他数据库支持,其生成的数据库文件能与流行统计软件 SAS、SPSS 和 Stata 无缝链接。2000 年开发团队公布了首个 EpiData 版本,并且成立了"EpiData 之友"来为 EpiData 开发提供资金支持。2002 年同时开发数据分析程序 Epic,旨在提供公共健康领域其它统计软件没有涉及的统计分

析，特别是“病例研究(cases studies)”，同年开发出 Epic 1.1 版本。EpiData 软件由世界各地志愿者翻译成各地版本，其中包括中文版本。

公司官方网址：http://www.epidata.dk/cn/index.htm

(三) EpiData 3.1 界面简介

EpiData 采用流行的线性流程操作，具体分为 6 步：①打开文件；②生成 REC 文件(问卷文件)；③建立 CHK 文件(检查文件)；④数据录入；⑤数据处理；⑥数据导出。第 3 步建立 CHK 文件为质量控制设置，第 5 步数据处理是为了进行双录入检测，这两个步骤可以省略，其他步骤一般都是必需的。



二、SPSS 简介

(一) 选择 SPSS 的理由

数据分析是科研的核心，一个没有数据支持的结论，甚至无法称之为科研结论，可以说数据分析是科研的核心竞争力，但遗憾的是，绝大多数科研人员却缺乏这种竞争力，在这个问题上，我国各医科大学的统计教学应当承担很大的责任，在临床研究生中教授 SAS 统计分析，如同教一个不会开车的人参加 F1 方程赛，结果可想而知了，以至最后，这些临床工作者们患有统计有用、统计难学的“统计恐惧症”。难道统计软件对普通临床医生来说只是一个只可远观的神物，不是的，那是因为没有选用 SPSS，选择 SPSS 18 中文版有如下理由：

1. 全中文的图形化界面，通俗易懂。虽然 SPSS 18 的中文版翻译和我们平常的统计术语略有差别，可能是由繁体中文版进化过来的，但是中文界面能帮助大家突破语言上的障碍。

2. 统计结果权威，在学术界和商业界，SPSS 和 SAS、STATA 并称为世界三大统计软件，其统计结果全球公认。

3. 统计方法丰富，SPSS 包括了如数据汇总、计数、交叉分析、分类、描述性统计分析、因子分析、回归和聚类分析等基本统计方法，以及线性回归模型和广义线性模型等高级统计方法，在科研、商业、农业、经济预测中发挥很大的作用。

(二) SPSS 的历史

1968 年，美国斯坦福大学三位专业背景截然不同的研究生 Norman H. Nie、C. Hadlai (Tex) Hull 和 Dale H. Bent 开发了最早的统计软件系统，称为“Statistical Package for the Social

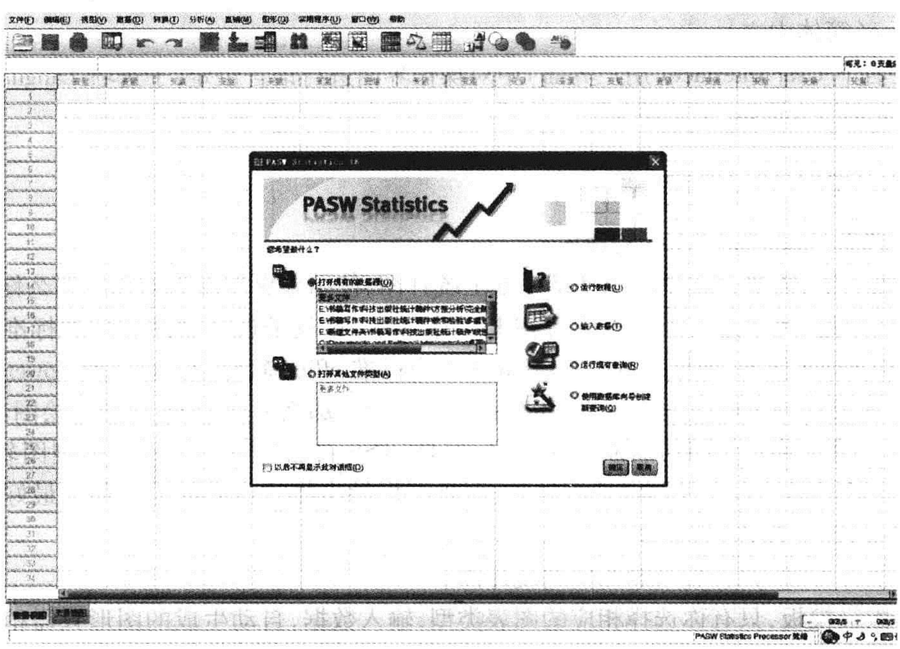
Sciences(社会科学统计软件包)",简称 SPSS,意图解决社会调查分析的统计问题。Nie 是一位社会科学家、斯坦福大学的博士生,在开发中承担了用户的角色,负责提出应用需求;Bent 是斯坦福大学的运筹学博士生,拥有数据分析方面的专业知识,负责设计 SPSS 系统文件结构;Hull 刚从斯坦福大学获得 MBA 学位,由他来负责编程。三人的工作得到了芝加哥大学的支持,但是后来 Bent 返回故乡加拿大,另外两位继续从事 SPSS 的研究和推广。1970 年麦格劳-希尔教育出版集团出版了第一本 SPSS 用户手册,从而引起了对该软件需求的迅猛增加。1975 年 SPSS 公司正式成立,1992 年正式在纳斯达克上市,2000 年随着 SPSS 产品服务领域的扩大和服务深度的增加,SPSS 公司正式将英文全称更改为"Statistical Product and Service Solutions(统计产品与服务解决方案)",而英文缩写 SPSS 没有改变。2009 年,随着 SPSS 公司成为预测分析领域的领导者,SPSS 公司将旗下的主要产品名称前面统一冠以 PASW 字样,全称为"Predictive Analysis Software",即"预测分析软件"。2010 年,随着 SPSS 公司被 IBM 收购,各子产品家族的前面不再冠以 PASW 为名,修改为统一加上 IBM SPSS 字样。现在公司旗下 4 个主要产品为:统计分析产品 SPSS statistics(原 SPSS),数据挖掘产品 SPSS Modeler(原 Clementine),数据收集产品 SPSS Data Collection family(原 Dimensions),企业应用服务 SPSS Collaboration and Deployment Services(原 Predictive Enterprise Services)。

公司官方网址:http://www. spss. com. cn

软件下载:公司官网提供 30 天无功能限制试用版本下载

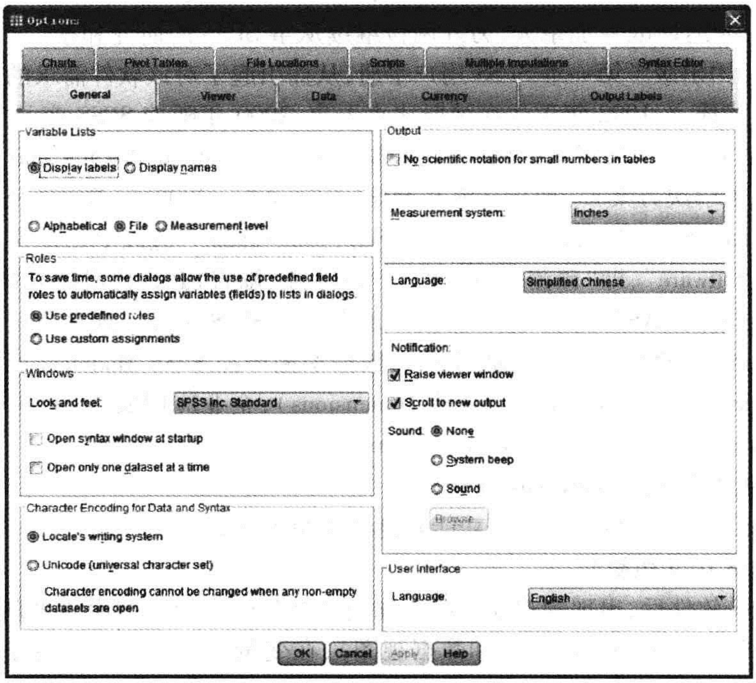
(三)SPSS 18 中文版简介

当打开 SPSS 软件后,出现的欢迎界面让你选择打开已有的数据库或者新建数据库,然后进入数据编辑窗口(又分为变量视图和数据视图),在变量视图窗口设置变量,然后在数据视图进行数据录入。在点击数据(D)下拉菜单选用相应窗口进行统计分析后,弹出结果窗口,显示统计分析结果。



(四)SPSS 18 英文版转换为中文版

SPSS 18 实际上为多语言版本,有时安装后默认为英文界面,我们需要在菜单栏 Edit(编辑)下拉菜单中 Options(选项)对话框中进行设定,如图所示:



1. 选择 General(常规),对 SPSS 整体进行设定。

2. 在右上 Output(输出)中对输出结果进行语言设定,子选项 Language(语言)选择 Simplified Chinese(简体中文)。

3. 在右下 User Interface(用户界面)中对操作界面进行语言设定,子选项 Language(语言)选择 Simplified Chinese(简体中文)。

三、GraphPad Prism 简介

(一)选择 GraphPad Prism 的理由

统计完成了,需要为重要的统计结果配上统计图,使结果更加直观可靠。那么你首先应当明确图形的用途,无非是出版发表或者在 PPT 中演示,这就是选择制图软件的标准。如果你立马想到用 Excel 作图,那你就 out 了。Excel 作为标准 office 软件,与 word 和 powerpoint 兼容性良好,但是要达到出版级发表水平,还存在一定差距。那么采用 SPSS 自带的绘图功能如何?虽然 SPSS 一直宣称它作图功能强大而丰富,但是 SPSS 自己却有一个专门的制图软件 Deltagraph,SPSS 对图形的自主控制灵活度较差,而且图形导出格式也有缺陷。那么最好的选择就是 GraphPad Prism 软件,选择该软件的理由如下:

1. 以统计方法为线索,绘制相应的统计图。

2. 出版级模板,只有你选择相应的图表类型,输入数据,自动生成的图形只需稍作修改即达到出版要求。

3. 灵活多变的图形修改方法,你可以在图形上任意绘制直线,添加文字,以满足特定的需求。

4. 简洁的排版功能,以符合出版要求。

5. 丰富的图形导出格式,可以导出成位图或矢量图,并且位图可以按照杂志要求,设定相应的 dpi。

6. 一键发布至 word 文档和 PPT 文档,与 office 文档无缝链接。

(二)GraphPad Prism 的历史

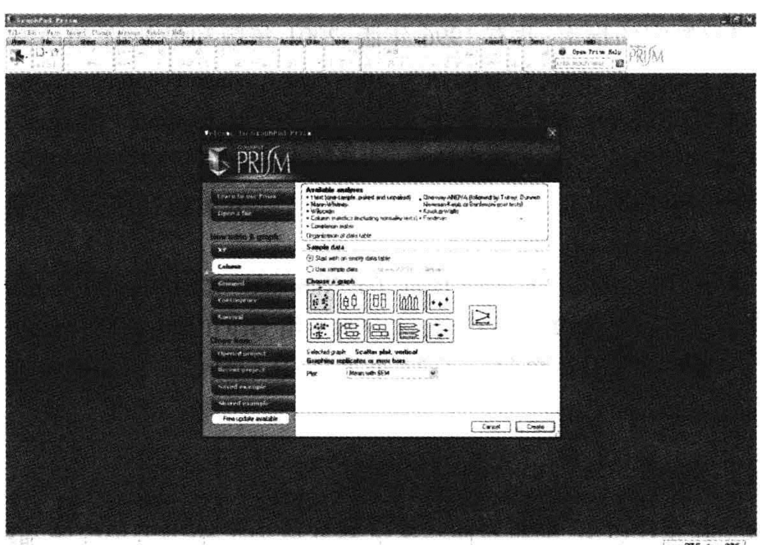
GraphPad 软件公司创始人 Harvey Motulsky 博士首先在加州大学药学院从事 α-和 β-肾上腺素受体对人体血细胞的调节作用研究,同时教授生物统计学。在这个过程中,他编写程序来分析自己的数据,同时将这些程序分享给他人使用。1984 年在科学出版社发布了"GraphPad"程序,一年后科学出版社决定停止销售该程序。Harvey Motulsky 博士于 1985 年在美国科学信息研究所(ISI)发表 GraphPad 程序的 2.0 版本,当一年后 ISI 决定停止销售该程序时,Harvey Motulsky 博士决定自己成立 GraphPad 软件公司来开发和销售程序。由于公司创始人的知识背景,所以 GraphPad 程序具有"科学家开发,为科学家服务"的鲜明特点,使用者不需要具有强大的统计知识,只需要关注自己的研究数据和研究目的,按照软件提示就可以迅速完成数据分析和图形绘制。最新的 GraphPad Prism 5 已经演变成集数据分析、图形绘制和回归分析于一体的综合性软件。该公司还发展出一系列产品,如简易的统计软件 GraphPad InStat,样本计算软件 GraphPad StatMate,在线免费统计软件 GraphPad Quick Calcs。现在世界上有超过 10 万人使用该软件,尤其在生物医学界具有很大的影响力。

公司官方网址:www. graphpad. com

软件下载:公司官网提供 30 天无功能限制试用版本下载

(三)GraphPad Prism 5 界面简介

软件开始的欢迎界面,在这个界面上,将对统计类别、数据输入格式和图形表达进行设定,也就是"模板",所以大大简化操作过程。



>>附录二 临床研究国际论文撰写指南

一个临床研究经历了文献查阅、研究设计、研究实施和数据分析步骤之后,终于快要达到了胜利的彼岸,也就是论文撰写和发表阶段。越来越多的中国医学研究人员在国际权威杂志上发表文章,更多的大学要求博士毕业时发表 SCI 文章,同时,由于语言障碍和对国际杂志规则的不熟悉,很多优秀的研究成果没能走出国门,刊登在高影响因子的专业杂志上,这不能不说是个很大的遗憾。要权威期刊接收我们的稿件,必须先了解它的游戏规则,这样才能在同一个语系中说话,才能更好地沟通和了解。在权威杂志方面,也认识到由于论文撰写质量不佳而埋没了优秀的研究成果,从 1996 年开始,国际权威杂志如《柳叶刀》、《美国医学会》编辑们联合起来,共同制定论文报告的撰写指南,要求研究人员在向杂志投稿时,按照其指南要求撰写,也可以将指南看成论文撰写的 SOP(Standard Operation Procedure,标准作业程序),研究人员只需要按照其详尽要求,完成对自己研究成果的报告即可。

如果将医学研究分为基于微观的实验室研究和基于宏观的人群研究,那么临床研究人员更多从事基于宏观的人群研究,而后者又可以按照研究设计的不同而分为临床试验性研究、临床观察性研究和系统评价三大类,合称为"临床三大研究",而对应的论文撰写指南为 CONSORT 声明、STROBE 声明和 PRISMA 声明,其各声明的核心为报告清单,即研究者在论文撰写时必须包括清单中的项目,而且各声明给出的说明文档例举了各清单项目报告的优秀范例并给出详尽解释:

1. CONSORT 声明(Consolidated Standards of Reporting Trials)　这是针对临床试验的论文撰写指南,也是最早制定的指南,共有 1996 年版、2001 年版和最新的 2010 年版,其清单包括 25 个项目,并给出了研究流程图,官方网站为 http:/www. consort-statement. org/。

2. STROBE 声明(The Strengthening the Reporting of Observational Studies in Epidemiology)　这是针对临床观察性研究的论文撰写指南,涉及了观察性研究的三种主要设计类型:队列研究、病例对照研究和横断面研究,于 2004 年开始制定,其清单包括了 22 个项目,官方网站为 http://www. strobe-statement. org/。

3. PRISMA 声明(Preferred Reporting Items for Systematic Reviews and Meta-Analyses)　这是针对系统评价和荟萃分析的论文撰写指南,该声明是由 QUOROM 声明(Quality of Reporting of Meta-Analyses)发展过来的,其清单包括了 27 个项目,还包括了一份研究流程图,官方网站为 http://www. prisma-statement. org/。

CONSORT 声明

自从 20 世纪 50 年代以来,随机对照临床试验就成为了探讨干预措施和药物效果的金标准,而 20 世纪 90 年代中期以来,临床试验报告的质量越来越受到重视,David Moher 和 Drum-

mond Rennie 于 1996 年组织制定了第一版 CONSORT 声明,并组成了 CONSORT 工作组,工作组包括了临床试验专家、统计学家、流行病学家和生物医学编辑,对 CONSORT 声明不断进行修订,形成了 CONSORT 声明 2001 版和 2010 版。声明得到了世界各地 400 多种期刊的明确支持,另外,国际医学期刊编辑委员会(International Committee of Medical Journal Editors, ICMJE, http://www.icmje.org)和科学编辑理事会(Council of Science Editors)和世界医学编辑协会(World Association of Medical Editors)也正式支持 CONSORT 声明,并且参与和支持的期刊和团体在不断壮大。CONSORT 2010 主要针对两组平行随机对照临床试验,这类试验占临床试验文献的半数以上,但声明中的绝大多数项目适用于各种类型的随机临床试验。CONSORT 2010 声明包括了一张含有 25 个项目的清单和一张流程图。

CONSORT 2010 清单

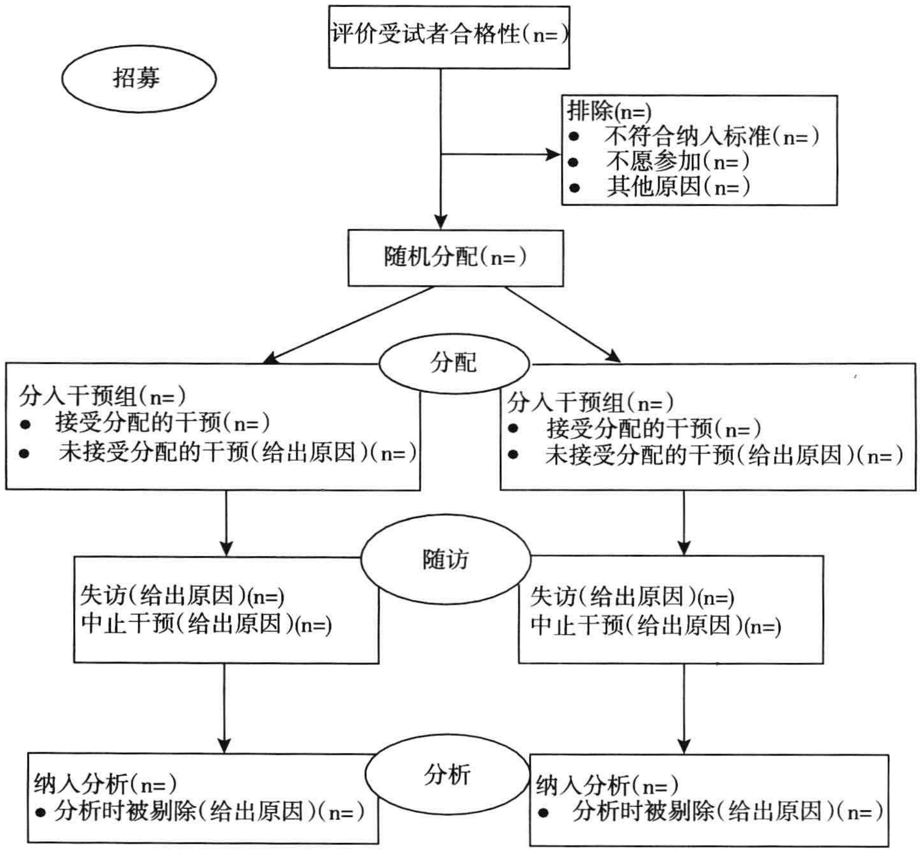
	条目	建议
题目和摘要		
	1a	文题能识别是随机临床试验
	1b	结构式摘要,包括试验设计、方法、结果、结论几个部分(具体的指导建议参见"CONSORT for abstracts")
引言		
背景和目的	2a	科学背景和对试验理由的解释
	2b	具体目的和假设
方法		
试验设计	3a	描述试验设计(诸如平行设计、析因设计),包括受试者分配入各组的比例
	3b	试验开始后对试验方法所作的重要改变(如合格受试者的挑选标准),并说明原因
受试者	4a	受试者合格标准
	4b	资料收集的场所和地点
干预措施	5	详细描述各组干预措施的细节以使他人能够重复,包括它们实际上是在何时、如何实施的
结局指标	6a	完整而确切地说明预先设定的主要和次要结局指标,包括它们是在何时、如何测评的
	6b	试验开始后对结局指标是否有任何更改,并说明原因
样本量	7a	如何确定样本量
	7b	必要时,解释中期分析和试验中止原则
随机方法:		
序列的产生	8a	产生随机分配序列的方法
	8b	随机方法的类型,任何限定的细节(如怎样分区组和各区组样本多少)

续表

	条目	建议
分配隐藏机制	9	用于执行随机分配序列的机制（例如按序编码的封藏法），描述干预措施分配之前为隐藏序列号所采取的步骤
实施	10	谁产生随机分配序列，谁招募受试者，谁给受试者分配干预措施
盲法	11a	如果实施了盲法，分配干预措施之后对谁设盲（例如受试者、医护提供者、结局评估者），以及盲法是如何实施的
	11b	如有必要，描述干预措施的相似之处
统计学方法	12a	用于比较各组主要和次要结局指标的统计学方法
	12b	附加分析的方法，诸如亚组分析和校正分析
结果		
受试者流程（极力推荐使用流程图）	13a	随机分配到各组的受试者例数，接受已分配治疗的例数，以及纳入主要结局分析的例数
	13b	随机分组后，各组脱落和被剔除的例数，并说明原因
招募受试者	14a	招募期和随访时间的长短，并说明具体日期
	14b	为什么试验中断或停止
基线资料	15	用一张表格列出每一组受试者的基线数据，包括人口学资料和临床特征
纳入分析的例数	16	各组纳入每一种分析的受试者数目（分母），以及是否按最初的分组分析
结局和估计值	17a	各组每一项主要和次要结局指标的结果，效应估计值及其精确性（如95%可信区间）
	17b	对于二分类结局，建议同时提供相对效应值和绝对效应值
辅助分析	18	所做的其他分析的结果，包括亚组分析和校正分析，指出哪些是预先设定的分析，哪些是新尝试的分析
危害	19	各组出现的所有严重危害或意外效果（具体的指导建议参见“CONSORT for harms”）
讨论		
局限性	20	试验的局限性，报告潜在偏倚和不精确的原因，以及出现多种分析结果的原因（如果有这种情况的话）
可推广性	21	试验结果被推广的可能性（外部可靠性，实用性）
解释	22	与结果相对应的解释，权衡试验结果的利弊，并且考虑其他相关证据
其他信息		
试验注册	23	临床试验注册号和注册机构名称
试验方案	24	如果有的话，在哪里可以获取完整的试验方案
资助	25	资助和其他支持（如提供药品）的来源，提供资助者所起的作用

摘自：周庆辉翻译，卞兆祥、刘建平审校. CONSORT 2010 声明：报告平行对照随机临床试验指南的更新. 中西医结合学报，2010，8(7)：604-612

CONSORT 2010 流程图



STROBE 声明

许多医学研究中的问题是采用观察性研究来探讨的。大多数对致病因素的研究都依赖于队列、病例对照或横断面研究。2004 年，由方法学家、研究人员和期刊编辑共同起草了一份观察性研究报告应纳入的条目清单，最后形成的 STROBE 声明包含了 22 个条目，分为论文的题目、摘要、引言、方法、结果和讨论等部分，其中 18 个条目适用于所有 3 种主要的观察性研究，其余 4 个条目则专门适用于队列、病例对照或横断面研究。

STROBE 声明清单

	条目	建议
题目和摘要		
	1	(a)在题目或摘要中用常用术语表明研究所采用的设计
		(b)在摘要中对所做工作和获得的结果作一个简明的总结
引言		
背景/原理	2	解释研究的科学背景和原理
	3	阐明具体研究目的,包括任何预先设定的假设
方法		
研究设计	4	尽早陈述研究设计的关键内容
研究设置	5	描述研究机构、研究地点及相关资料,包括招募的时间范围,暴露、随访和数据收集等
参与者	6	(a)队列设计——描述纳入标准、参与者来源和选择方法,随访方法
		病例对照设计——描述纳入标准、病例和对照的来源及确认病例和选择对照的方法,病例和对照选择的原理
		横断面设计——描述纳入标准,参与者的来源和选择方法
		(b)队列设计——对于配对设计,应说明配对标准及暴露和非暴露的人数
		病例对照设计——对于配对设计,应说明配对标准和每个病例配对的对照数
变量	7	明确定义结局、暴露、预测因子,可能的混杂因素及效应修饰因素,如果相关,给出诊断标准
数据来源/测量	8	将每个有意义的变量,给出数据来源和详细的测量方法。如果有一个以上的组,描述各组之间测量方法的可比性
偏倚	9	描述解决潜在偏倚的方法
样本大小	10	描述样本量的确定方法
定量变量	11	解释定量变量是如何分析的。如果相关,描述各组的方法和原因
统计方法	12	(a)描述所用的所有统计方法,包括减少混杂因素的方法
		(b)描述所有分析亚组和交互作用的方法
		(c)解释如何解决数据缺失
		(d)队列设计-如果相关,描述解决失访问题的方法
		病例对照设计-如果相关,描述如何对病例和对照进行配对
		横断面设计-如果相关,描述考虑到抽样策略的分析方法
		(e) 描述所用的灵敏性分析方法

续表

	条目	建议
结果		
参与者	13	(a) 报告研究各阶段参与者的人数,如可能合格的人数,参与合格性检查的人数,证实合格的人数,纳入研究的人数,完成随访的人数及完成分析的人数
		(b)解释在各阶段参与者推出研究的原因
		(c)考虑使用流程图
描述性数据	14	(a)描述参与者的特征(如人口统计学,临床和社会特征)以及暴露和潜在混杂因素的相关信息
		(b)描述就每一个待测变量而言缺失数据的参与者人数
		(c)队列设计-总结随访时间(如平均随访时间和全部随访时间)
结局数据	15	队列设计-报告随时间变化的结局事件数或综合指标
		病例对照设计-报告各种暴露类别的人数或暴露综合指标
		横断面设计-报告结局事件数或综合指标
主要结果	16	(a)报告未校正的估计值。如果相关,给出混杂因素校正后的估计值及其精确度(如95%可信区间)。说明按照哪些混杂因素进行了校正以及选择这些因素进行校正的原因
		(b)如对连续变量进行分组,要报告每组观察值的范围
		(c)对有意义的危险因素,最好把相对危险转换成针对有意义的时间范围的绝对危险度
其他分析	17	报告进行过的其他分析,如亚组分析、交互作用分析和灵敏性分析
讨论		
关键结果	18	根据研究目标概括关键结果
局限性	19	讨论研究的局限性,包括潜在偏倚或不准确的来源,讨论任何潜在偏倚的方向和大小
解释	20	结合研究目标、研究局限性、多重分析、相似研究的结果和其他相关证据,谨慎给出一个总体的结果解释
可推广性	21	讨论研究结果的普适性(外推有效性)
其他信息		
资金来源	22	提供研究资金的来源和资助机构在研究中的作用。如果相关,提供资助机构在本文基于的初始研究中作用

摘自:赵乐翻译,邹光勇、杜怀东审校. 加强流行病学中观察性研究报告质量(STROBE)声明:观察性研究报告规范. 世界临床医学, 2008,2(1):78-82

PRISMA 声明

系统综述和荟萃分析(又称 meta 分析)在医疗保健领域已变得越来越重要,为了提高其报

告质量,1996 年,一个国际性小组发布了 meta 分析报告质量(Quality of Reporting of Meta-Analyses, QUOROM)指南,该指南着重关注随机对照试验 meta 分析的报告。2005 年 6 月,由综述作者、方法学家、临床医师、医学编辑和使用者组成的小组在加拿大着手修订和扩展 QUOROM 清单和流程图,并且将其更名为系统综述和 meta 分析优先报告的条目(Preferred Reporting Items for Systematic Reviews and Meta-Analyses, PRISMA)。PRISMA 声明包括了一张含有 27 个条目组成的清单和一个四阶段的流程图。

PRISMA 声明清单

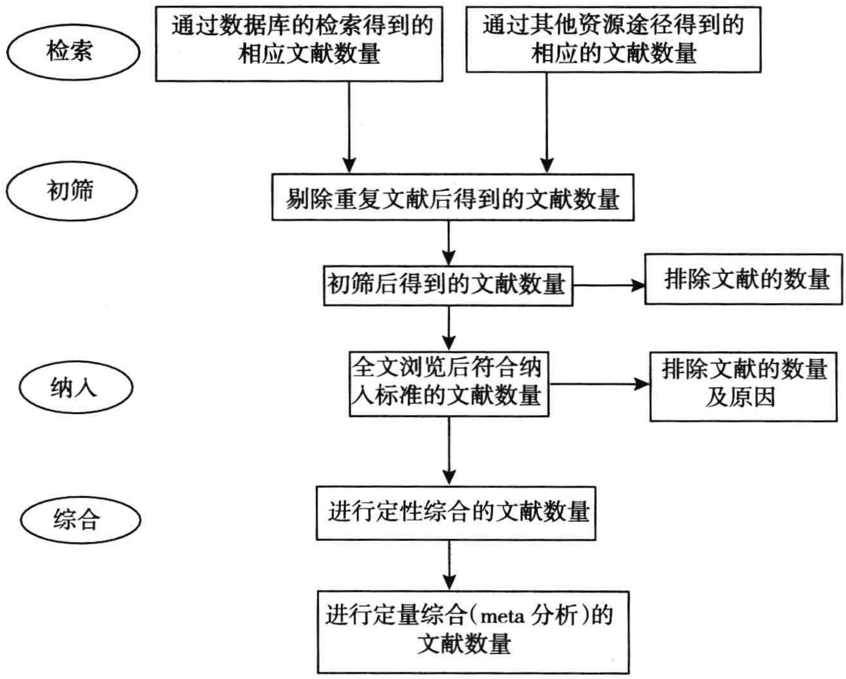
	条目	清单
标题		
标题	1	明确本研究报告是系统综述、meta 分析,还是两者兼有
摘要		
结构式摘要	2	提供结构式摘要包括背景、目的、资料来源、纳入研究的标准、研究对象和干预措施、研究评价和综合的方法、结果、局限性、结论和主要发现、系统综述的注册号
前言		
理论基础	3	介绍当前已知的研究理论基础
目的	4	通过对研究对象、干预措施、对照措施、结局指标和研究类型(participants, interventions, comparisons, outcomes, study design, PICOS)5 个方面为导向的问题提出所需要解决的清晰明确的研究问题
方法		
方案和注册	5	如果已有研究方案,则说明方案内容并给出可获得该方案的途径(如网址),并且提供现有的已注册的研究信息,包括注册号
纳入标准	6	将指定的研究特征(如 PICOS 和随访的期限)和报告的特征(如检索年限、语种和发表情况)作为纳入研究的标准,并给出合理的说明
信息来源	7	针对每次检索及最终检索的结果描述所有文献信息的来源(如资料库文献,与研究作者联系获取相应的文献)
检索	8	至少说明一个资料库的检索方法,包括所有的检索策略的使用,使得检索结果可以重现
研究选择	9	说明纳入研究被选择的过程(包括初筛、合格性鉴定及纳入系统综述等步骤,据实还可包括纳入 meta 分析的过程)
资料提取	10	描述资料提取的方法(例如预提取表格、独立提取、重复提取)以及任何向报告作者获取或确认资料的过程
资料条目	11	列出并说明所有资料相关的条目(如 PICOS 和资金来源),以及作出的任何推断和简化形式
单个研究存在的偏倚	12	描述用于评价单个研究偏倚的方法(包括该方法是否用于研究层面或结局层面),以及在资料综合种该信息如何被利用

续表

	条目	清单
概括效应指标	13	说明主要的综合结局指标，如危险度比值（risk ratio）、均数差（difference in means）
结果综合	14	描述结果综合的方法，如果进行了 meta 分析，则说明异质性检验的方法
研究偏倚	15	详细评估可能影响数据综合结果的可能存在的偏倚（如发表偏倚和研究中的选择性报告偏倚）
其他分析	16	对研究中其他的分析方法进行描述（如敏感性分析或亚组分析，meta 回归分析），并说明哪些分析方法是预先制订的
结果		
研究选择	17	报告初筛的文献数，评价符合纳入标准的文献数以及最终纳入研究的文献数，同时给出每一步排除文献的原因，最好提供流程图
研究特征	18	说明每一个被提取资料的文献的特征（如样本含量、PICOS 和随访时间）并提供引文出处
研究内部偏倚风险	19	说明每个研究中可能存在偏倚的相关数据，如果条件允许，还需要说明结局层面的评估（见条目 12）
单个研究的结果	20	针对所有结局指标（有效性或有害性），说明每个研究的各干预组结果的简单合并（a），以及综合效应值及其可信区间（b），最好以森林图形式报告
结果的综合	21	说明每个 meta 分析的结果，包括可信区间和异质性检验的结果
研究间偏倚	22	说明研究间可能存在偏倚的评价结果（见条目 15）
其他分析	23	如果有，给出其他分析的结果（如敏感性分析或亚组分析，meta 回归分析，见条目 16）
讨论		
证据总结	24	总结研究的主要发现，包括每一个主要结局的证据强度；分析它们与主要利益集团的关联性（如医疗保健的提供者、使用者及政策决策者）
局限性	25	探讨研究层面和结局层面的局限性（如偏倚的风险），以及系统综述的局限性（如检索不全面，报告偏倚等）
结论	26	给出对结果的概要性的解析，并提出对未来研究的提示
资金支持		
资金	27	描述本系统综述的资金来源和其他支持（如提供资料）以及资助者在完成系统综述中所起的作用

摘自：李迅、曹卉娟翻译，刘建平审校. 系统综述和荟萃分析优先报告的条目：PRISMA 声明. 中西医结合学报，2009，7（9）：889-896

PRISMA 流程图


检索
通过数据库的检索得到的相应文献数量
通过其他资源途径得到的相应的文献数量
初筛
剔除重复文献后得到的文献数量
初筛后得到的文献数量
排除文献的数量
纳入
全文浏览后符合纳入标准的文献数量
排除文献的数量及原因
综合
进行定性综合的文献数量
进行定量综合(meta 分析)的文献数量

> > 参考文献

[1] Agresti A. An introduction to categorical data analysis. New York: Wiley-Inter-science Publication, 1996.

[2] Beth DS, et Robert GT. Basic Clinical Biostatistics. 2nd ed. Appleton and Lange, 1994.

[3] C. Li. Introduction to Experimental Statistics. New York: McGraw-Hill Book Company, 1964.

[4] Chinns S. The assessment of methods of measurement. Statistics in Medicine, 1990,9:351.

[5] Crower MJ, Hand DJ. Analysis of repeated measures. London: Chapman and Hall, 1990.

[6] Fleiss JL. Statistical Methods for Rates and Proportions. 2nd ed. John Wiley & Sons, 1981.

[7] Hosmer DW Jr, Listicemeshow S. Applied Logistic Regression. New York: John Wiley & Sons, 1989.

[8] Kendall M, Gibons JD. Rank correlation methods. 5th ed. New York: Oxford University Press, 1990.

[9] Lawless JF. Statistical Models and Methods for Lifetime Data. New York: John Wiley & Sons, 1982.

[10] Lee ET. Statistical Methods for survival Data Analysis. New York: John Wiley & Sons, 1992.

[11] Maurice Kendall, Jean Dickinon Gibbons. Rank correlation methods. 4th ed. London: Edword Arnold, 1990.

[12] Rosner B. Foundations of Biostatistics. Harvard university Duxburg Press, 1982.

[13] SPSS Inc. SPSS Advanced Model 18.0. USA 2007.

[14] SPSS Inc. SPSS BASE 18.0. USA 2007.

[15] Vonesh EF, Chinchilli VM. Linear and nonlinear models for the analysis of repeated measurements. New York: Marcel Deker Inc, 1997.

[16] Harvey J. Motulsky. Intuitive Biostatistics. 2nd ed. Oxford University Press, 2010.

[17] GraphPad Software Inc. GraphPad Prism help Document. 2009.

[18] Kleinbaum DG. (英文版)应用回归分析和其他多元方法[M].第3版. 北京：机械工业出版社, 2003.

[19] 曹素华. 实用医学多因素统计方法[M]. 上海：上海医科大学出版社, 1998.

[20] 陈峰.应用多元统计分析方法[M]. 北京：中国统计出版社, 2001.

[21] 方积乾, 陆盈. 现代医学统计学[M]. 北京：人民卫生出版社, 2002.

[22] 方积乾. 卫生统计学[M]. 第6版. 北京：人民卫生出版社, 2006.

[23] 胡良平. 现代统计学与SAS应用[M]. 北京：军事医学科学出版社, 1996.

[24] 蒋知俭. 医学统计学[M]. 北京：人民卫生出版社, 1997.

[25] 谭红专. 现代流行病学[M]. 北京：人民卫生出版社, 2001.

[26] 孙振球. 医学统计学[M]. 第2版. 北京：人民卫生出版社, 2008.

[27] 张文彤, 闫洁. SPSS统计分析基础教程[M]. 北京：高等教育出版社, 2004.

[28] 张文彤. SPSS统计分析高级教程[M]. 北京：高等教育出版社, 2004.

[29] 宇传华, 颜杰. Excel与数据分析[M]. 北京：电子工业出版社, 2002.

[30] 许军. EpiData 3.02数据管理软件实用教程[M]. 北京：军事医学科学出版社, 2006.